

# STABLE COMPLETE INTERSECTIONS

LORENZO ROBBIANO AND MARIA LAURA TORRENTE

**ABSTRACT.** A complete intersection of  $n$  polynomials in  $n$  indeterminates has only a finite number of zeros. In this paper we address the following question: how do the zeros change when the coefficients of the polynomials are perturbed? In the first part we show how to construct semi-algebraic sets in the parameter space over which all the complete intersection ideals share the same number of isolated real zeros. In the second part we show how to modify the complete intersection and get a new one which generates the same ideal but whose real zeros are more stable with respect to perturbations of the coefficients.

## 1. INTRODUCTION

What is the defining (or vanishing) ideal of a finite set  $\mathbb{X}$  of points in the affine space? The standard answer is that it is the set of all the polynomials which vanish at  $\mathbb{X}$ . And there are very efficient methods to compute it, based on Buchberger-Möller's algorithm (see for instance [1], [2] and [6]).

However, the logical and computational environment changes completely when the coordinates of the points are perturbed by errors, a situation which is normal when dealing with real world problems. In that case one has to use approximation and to consider the question of stability. Introductory material about this topic can be found in the book [3], in particular in the paper [14] and its bibliography.

The methods used so far share the strategy of modifying the Buchberger-Möller Algorithm and compute a Gröbner basis or a border basis of an ideal of polynomials which *almost vanish* at  $\mathbb{X}$  (see for instance [10] and [11]). A key remark is that, whatever algorithm is used, at a certain moment one has computed  $n$  polynomials  $f_1, \dots, f_n$  which generate a zero-dimensional ideal. Since the dimension has dropped from  $n$  to zero, the  $n$  polynomials form a complete intersection which almost vanishes at  $\mathbb{X}$ . Further steps in the algorithm will be used to eliminate spurious points and to produce a Gröbner or border basis.

Now, a complete intersection of  $n$  polynomials in  $n$  indeterminates has only a finite number of zeros, and the main question is: how do the zeros change when the coefficients of the polynomials are perturbed? Can we devise a strategy to make the situation reasonably stable? In other words, can we change the generating polynomials so that the stability of their common zeros increases? It is well-known that for a linear system with  $n$  equations and  $n$  unknowns, the most stable situation occurs when the coefficient matrix is orthonormal. Is there an analogue to orthonormality when we deal with polynomial systems?

---

*Date:* January 12, 2013.

*2010 Mathematics Subject Classification.* Primary 13C40, Secondary 14M10, 65F35, 65H04.

*Key words and phrases.* complete intersection, condition number.

In numerical analysis the condition number of a problem measures the sensitivity of the solution to small changes in the input data, and so it reveals how numerically well-conditioned the problem is. There exist a huge body of results about condition numbers for various numerical problems, for instance the solution of a linear system, the problem of matrix inversion, the least squares problem, and the computation of eigenvalues and eigenvectors.

On the other hand, not very much is known about condition numbers of polynomial systems. As a notable exception we mention the paper [17] of Shub and Smale who treated the case of zero-dimensional homogeneous polynomial systems; later on their result was extended by Dégot (see [8]) to the case of positive-dimensional homogeneous polynomial systems.

Tackling the above mentioned problem entails a preliminary analysis of the following question of algebraic nature. If we are given a zero-dimensional complete intersection of polynomials with simple zeros, how far can we perturb the coefficients so that the zeros remain smooth and their number does not change? It is quite clear that smoothness and constancy of the number of zeros are essential if we want to consider the perturbation a good one.

Starting with the classical idea that a perturbed system is a member of a family of systems, we describe a good subset of the parameter space over which the members of the family share the property that their zero sets have the same number of smooth real points. This is the content of Section 2 where we describe a free (see Proposition 2.6), and a smooth (see Theorem 2.12) locus in the parameter space. Then we provide a suitable algorithm to compute what we call an  $I$ -optimal subscheme of the parameter space (see Corollary 2.16): it is a subscheme over which the complete intersection schemes are smooth and have the same number of complex points. The last important result of Section 2 is Theorem 2.20 which proves the existence of an open non-empty semi-algebraic subscheme of the  $I$ -optimal subscheme over which the number of real zeros is constant.

Having described a good subscheme of the parameter space over which we are allowed to move, and hence over which we can perturb our data, we pass in Section 3 to the next problem and concentrate our investigation on a single point of the zero set. After some preparatory results, we introduce a local condition number (see Definition 3.14) and with its help we prove Theorem 3.15 which has the merit of fully generalizing a classical result in numerical linear algebra (see Remark 3.16).

The subsequent short Section 4 illustrates how to manipulate the equations in order to lower, and sometimes to minimize, the local condition number (see Proposition 4.1). Then we concentrate on the case of the matrix 2-norm and show how to achieve the minimum when the polynomials involved have equal degree (see Proposition 4.3). The final Section 5 describes examples which indicate that our approach is good, in particular we see that when the local condition number is lowered, indeed the corresponding solution is more stable.

This paper reports on the first part of a wider investigation. Another paper is already planned to describe how to deal with global condition numbers and how to generalize our method to the case where the polynomials involved have arbitrary degrees.

All the supporting computations were performed with CoCoA (see [7]). We thank Marie-Françoise Roy and Saugata Basu for some help in the proof of Theorem 2.20.

## 2. FAMILIES OF COMPLETE INTERSECTIONS

Given a zero-dimensional smooth complete intersection  $\mathbb{X}$ , we want to embed it into a family of zero-dimensional complete intersections and study when and how  $\mathbb{X}$  can move inside the family. In particular, we study the locus of the parameter-space over which the fibers are smooth with the same number of points as  $\mathbb{X}$ , and we give special emphasis to the case of real points.

We start the section by recalling some definitions. The notation is borrowed from [15] and [16], in particular we let  $x_1, \dots, x_n$  be indeterminates and let  $\mathbb{T}^n$  be the monoid of the power products in the symbols  $x_1, \dots, x_n$ . Most of the times, for simplicity we use the notation  $\mathbf{x} = x_1, \dots, x_n$ . If  $K$  is a field, the multivariate polynomial ring  $K[\mathbf{x}] = K[x_1, \dots, x_n]$  is denoted by  $P$ , and if  $f_1(\mathbf{x}), \dots, f_k(\mathbf{x})$  are polynomials in  $P$ , the set  $\{f_1(\mathbf{x}), \dots, f_k(\mathbf{x})\}$  is denoted by  $\mathbf{f}(\mathbf{x})$  (or simply by  $\mathbf{f}$ ). Finally, we denote the *polynomial system* associated to  $\mathbf{f}(\mathbf{x})$  by  $\mathbf{f}(\mathbf{x}) = 0$  (or simply by  $\mathbf{f} = 0$ ), and we say that the system is zero-dimensional if the ideal generated by  $\mathbf{f}(\mathbf{x})$  is zero-dimensional (see [15], Section 3.7).

Easy examples show that, unlike the homogeneous case, in the inhomogeneous case regular sequences are not independent of the order of their entries. For instance, if  $f_1 = y(x+1)$ ,  $f_2 = z(x+1)$ ,  $f_3 = x$ , then  $(f_1, f_2, f_3)$  is not a regular sequence, while  $(f_3, f_1, f_2)$  is such. However, we prefer to avoid a distinction between these cases, and we call them *complete intersections*. In other words, we use the following definition.

**Definition 2.1.** Let  $t$  be a positive integer, let  $\mathbf{f}(\mathbf{x})$  be a set of  $t$  polynomials in  $P = K[x_1, \dots, x_n]$  and let  $I$  be the ideal generated by  $\mathbf{f}(\mathbf{x})$ .

- (a) The set  $\mathbf{f}(\mathbf{x})$  (and the ideal  $I$ ) is called a **complete intersection** if the equality  $\dim(P/I) = n - t$  holds.
- (b) The set  $\mathbf{f}(\mathbf{x})$  (and the ideal  $I$ ) is called a **zero-dimensional complete intersection** if it is a complete intersection and  $t = n$ .

Let  $n$  be a positive integer, let  $P$  denote the polynomial ring  $K[x_1, \dots, x_n]$ , let  $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$  be a zero-dimensional complete intersection, and let  $I$  be the ideal of  $P$  generated by  $\mathbf{f}(\mathbf{x})$ . We let  $m$  be a positive integer and let  $\mathbf{a} = (a_1, \dots, a_m)$  be an  $m$ -tuple of indeterminates which will play the role of parameters. If  $F_1(\mathbf{a}, \mathbf{x}), \dots, F_n(\mathbf{a}, \mathbf{x})$  are polynomials in  $K[\mathbf{a}, \mathbf{x}]$  we let  $F(\mathbf{a}, \mathbf{x}) = 0$  be the corresponding family of systems of equations parametrized by  $\mathbf{a}$ , and the ideal generated by  $F(\mathbf{a}, \mathbf{x})$  in  $K[\mathbf{a}, \mathbf{x}]$  is denoted by  $I(\mathbf{a}, \mathbf{x})$ . If the scheme of the  $\mathbf{a}$ -parameters is  $\mathcal{S}$ , then there is a  $K$ -algebra homomorphism  $\varphi : K[\mathbf{a}] \rightarrow K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x})$  or, equivalently, a morphism of schemes  $\Phi : \text{Spec}(K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x})) \rightarrow \mathcal{S}$ .

Although it is not strictly necessary for the theory, for our applications it suffices to consider independent parameters. Here is the formal definition.

**Definition 2.2.** If  $\mathcal{S} = \mathbb{A}_K^m$  and  $I(\mathbf{a}, \mathbf{x}) \cap K[\mathbf{a}] = (0)$ , then the parameters  $\mathbf{a}$  are said to be **independent** with respect to  $F(\mathbf{a}, \mathbf{x})$ , or simply independent if the context is clear.

The first important step is to embed the system  $\mathbf{f}(\mathbf{x}) = 0$  into a family, but we must be careful and exclude families of the following type.

**Example 2.3.** Consider the family  $F(a, \mathbf{x}) = \{x_1(ax_2 + 1), x_2(ax_2 + 1)\}$ . It is a zero dimensional complete intersection only for  $a = 0$  while the generic member is positive-dimensional.

**Definition 2.4.** Let  $\mathbf{f}(\mathbf{x})$  be a set of polynomials in  $P = K[x_1, \dots, x_n]$  so that  $\mathbf{f}(\mathbf{x})$  is a zero-dimensional complete intersection and let  $F(\mathbf{a}, \mathbf{x})$  be a family parametrized by  $m$  independent parameters  $\mathbf{a}$ . We say that  $F(\mathbf{a}, \mathbf{x})$  (and similarly  $K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x})$  and  $\text{Spec}(K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x}))$ ) is a **generically zero-dimensional family containing  $\mathbf{f}(\mathbf{x})$** , if  $\mathbf{f}(\mathbf{x})$  is a member of the family and the generic member of the family is a zero-dimensional complete intersection.

A theorem called *generic flatness* (see [9], Theorem 14.4) prescribes the existence of a non-empty Zariski-open subscheme  $\mathcal{U}$  of  $\mathcal{S}$  over which the morphism  $\Phi^{-1}(\mathcal{U}) \rightarrow \mathcal{U}$  is *flat*. In particular, it is possible to explicitly compute a subscheme over which the morphism is free. To do this, Gröbner bases reveal themselves as a fundamental tool.

**Definition 2.5.** Let  $F(\mathbf{a}, \mathbf{x})$  be a generically zero-dimensional family which contains a zero-dimensional complete intersection  $\mathbf{f}(\mathbf{x})$ . Let  $\mathcal{S} = \mathbb{A}_K^m$  be the scheme of the independent  $\mathbf{a}$ -parameters and let  $\Phi : \text{Spec}(K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x})) \rightarrow \mathcal{S}$  be the associated morphism of schemes. A dense Zariski-open subscheme  $\mathcal{U}$  of  $\mathcal{S}$  such that  $\Phi^{-1}(\mathcal{U}) \rightarrow \mathcal{U}$  is free (flat, faithfully flat), is said to be an  *$I$ -free* ( *$I$ -flat*,  *$I$ -faithfully flat*) subscheme of  $\mathcal{S}$  or simply an  *$I$ -free* ( *$I$ -flat*,  *$I$ -faithfully flat*) scheme.

**Proposition 2.6.** *With the above assumptions and notation, let  $I(\mathbf{a}, \mathbf{x})$  be the ideal generated by  $F(\mathbf{a}, \mathbf{x})$  in  $K[\mathbf{a}, \mathbf{x}]$ , let  $\sigma$  be a term ordering on  $\mathbb{T}^n$ , let  $G(\mathbf{a}, \mathbf{x})$  be the reduced  $\sigma$ -Gröbner basis of the ideal  $I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}]$ , let  $d(\mathbf{a})$  be the least common multiple of all the denominators of the coefficients of the polynomials in  $G(\mathbf{a}, \mathbf{x})$ , and let  $T = \mathbb{T}^n \setminus \text{LT}_\sigma(I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}])$ .*

- (a) *The open subscheme  $\mathcal{U}$  of  $\mathbb{A}_K^m$  defined by  $d(\mathbf{a}) \neq 0$  is  $I$ -free.*
- (b) *The multiplicity of each fiber over  $\mathcal{U}$  coincides with the cardinality of  $T$ .*

*Proof.* The assumption that  $F(\mathbf{a}, \mathbf{x})$  is a generically zero-dimensional family implies that  $\text{Spec}(K(\mathbf{a})[\mathbf{x}]/I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}]) \rightarrow \text{Spec}(K(\mathbf{a}))$  is finite, in other words that  $K(\mathbf{a})[\mathbf{x}]/I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}]$  is a finite-dimensional  $K(\mathbf{a})$ -vector space. A standard result in Gröbner basis theory (see for instance [15], Theorem 1.5.7) shows that the residue classes of the elements in  $T$  form a  $K(\mathbf{a})$ -basis of this vector space. We denote by  $\mathcal{U}$  the open subscheme of  $\mathbb{A}_K^m$  defined by  $d(\mathbf{a}) \neq 0$ . For every point in  $\mathcal{U}$ , the given reduced Gröbner basis evaluates to the reduced Gröbner basis of the corresponding ideal. Therefore the leading term ideal is the same for all these fibers, and so is its complement  $T$ . If we denote by  $K[\mathbf{a}]_{d(\mathbf{a})}$  the localization of  $K[\mathbf{a}]$  at the element  $d(\mathbf{a})$  and by  $I(\mathbf{a}, \mathbf{x})^e$  the extension of the ideal  $I(\mathbf{a}, \mathbf{x})$  to the ring  $K[\mathbf{a}]_{d(\mathbf{a})}$ , then  $K[\mathbf{a}]_{d(\mathbf{a})}[\mathbf{x}]/I(\mathbf{a}, \mathbf{x})^e$  turns out to be a free  $K[\mathbf{a}]_{d(\mathbf{a})}$ -module. So claim (a) is proved. Claim (b) follows immediately from (a).  $\square$

**Remark 2.7.** We collect here a few remarks about this proposition. First of all, we observe that the term ordering  $\sigma$  can be chosen arbitrarily. Secondly, for every  $\alpha \in \mathcal{U}$  let  $L_\alpha$  be the leading term ideal of the corresponding ideal  $I_\alpha$ . If  $\sigma$  is a degree-compatible term ordering, then  $L_\alpha$  is also the leading term ideal of the homogenization  $I_\alpha^{\text{hom}}$  of  $I_\alpha$  (see [16], Proposition 5.6.3 and its proof).

**Example 2.8.** We consider the ideal  $I = (f_1, g)$  of  $K[x, y]$  where  $f_1 = x^3 - y$ ,  $g = x(x-1)(x+1)(x-2)(x+2)(x-3)(x+3)(x+13)(x^2+x+1)$ . We check that  $I = (f_1, f_2)$  where  $f_2 = xy^3 + 504x^2y - 183xy^2 + 14y^3 - 504x^2 + 650xy - 147y^2 -$

$468x + 133y$ . It is a zero-dimensional complete intersection and we embed it into the family  $I(\mathbf{a}, \mathbf{x}) = (ax^3 - y, g)$ . If we pick  $\sigma = \text{Lex}$  with  $y > x$  and perform the computation as suggested by the proposition, we get the freeness of the family for all  $a$ . Instead, we get the freeness of the family  $I(\mathbf{a}, \mathbf{x}) = (ax^3 - y, f_2)$  for  $a \neq 0$  (see a further discussion in Example 2.14).

**Example 2.9.** We let  $P = \mathbb{C}[x]$ , the univariate polynomial ring, and embed the ideal  $I$  generated by the following polynomial  $x^2 - 3x + 2$  into the generically zero-dimensional family  $F(\mathbf{a}, x) = \{a_1x^2 - a_2x + a_3\}$ . Such family is given by the canonical  $K$ -algebra homomorphism

$$\varphi : \mathbb{C}[\mathbf{a}] \longrightarrow \mathbb{C}[\mathbf{a}, x]/(a_1, a_2, a_3)/(a_1x^2 - a_2x + a_3)$$

It is a zero dimensional complete intersection for

$$\{\alpha \in \mathbb{C}^3 \mid \alpha_1 \neq 0\} \cup \{\alpha \in \mathbb{C}^3 \mid \alpha_1 = 0, \alpha_2 \neq 0\}.$$

It represents two distinct smooth points for

$$\{\alpha \in \mathbb{C}^3 \mid \alpha_1 \neq 0, \alpha_2^2 - 4\alpha_1\alpha_3 \neq 0\}.$$

It represents a smooth point for  $\{\alpha \in \mathbb{C}^3 \mid \alpha_1 = 0, \alpha_2 \neq 0\}$ .

It is not a zero-dimensional complete intersection for  $\{\alpha \in \mathbb{C}^3 \mid \alpha_1 = 0, \alpha_2 = 0\}$ .

This kind of examples motivates the following definition.

**Definition 2.10.** Let  $F(\mathbf{a}, \mathbf{x})$  be a generically zero-dimensional family containing a zero-dimensional complete intersection  $\mathbf{f}(\mathbf{x})$ . Let  $\mathcal{S} = \mathbb{A}_K^m$  be the scheme of the independent  $\mathbf{a}$ -parameters and let  $\Phi : \text{Spec}(K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x})) \longrightarrow \mathcal{S}$  be the associated morphism of schemes. A dense Zariski-open subscheme  $\mathcal{U}$  of  $\mathcal{S}$  such that  $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$  is **smooth**, i.e. all the fibers of  $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$  are zero-dimensional smooth complete intersections, is said to be an  **$I$ -smooth** subscheme of  $\mathcal{S}$  or simply an  **$I$ -smooth scheme**.

For instance in Example 2.9 we have the equality  $\mathcal{S} = \mathbb{A}_{\mathbb{C}}^3$  and the following open set  $\mathcal{U} = \{\alpha \in \mathbb{C}^3 \mid \alpha_1 \neq 0, \alpha_2^2 - 4\alpha_1\alpha_3 \neq 0\}$  is  $I$ -smooth.

**Remark 2.11.** We observe that a dense  $I$ -smooth scheme may not exist. It suffices to consider the ideal  $I = (x - 1)^2$  embedded into the family  $(x - a)^2$ . In any event, a practical way to find one, if there is one, is via Jacobians, as we are going to show.

**Theorem 2.12.** Let  $F(\mathbf{a}, \mathbf{x})$  be a generically zero-dimensional family containing a zero-dimensional complete intersection  $\mathbf{f}(\mathbf{x})$ . We let  $\mathcal{S} = \mathbb{A}_K^m$  be the scheme of the independent  $\mathbf{a}$ -parameters, let  $I(\mathbf{a}, \mathbf{x})$  be the ideal generated by  $F(\mathbf{a}, \mathbf{x})$  in  $K[\mathbf{a}, \mathbf{x}]$ , let  $D(\mathbf{a}, \mathbf{x}) = \det(\text{Jac}_F(\mathbf{a}, \mathbf{x}))$  be the determinant of the Jacobian matrix of  $F(\mathbf{a}, \mathbf{x})$  with respect to the indeterminates  $\mathbf{x}$ , let  $J(\mathbf{a}, \mathbf{x})$  be the ideal sum  $I(\mathbf{a}, \mathbf{x}) + (D(\mathbf{a}, \mathbf{x}))$  in  $K[\mathbf{a}, \mathbf{x}]$ , and let  $H$  be the ideal in  $K[\mathbf{a}]$  defined by the equality  $H = J(\mathbf{a}, \mathbf{x}) \cap K[\mathbf{a}]$ .

- (a) There exists an  $I$ -smooth subscheme of  $\mathcal{S}$  if and only if  $H \neq (0)$ .
- (b) If  $0 \neq h(\mathbf{a}) \in H$  then the open subscheme of  $\mathcal{S}$  defined by the inequality  $h(\mathbf{a}) \neq 0$  is  $I$ -smooth.

*Proof.* To prove one implication of claim (a), and simultaneously claim (b), we assume that  $H \neq (0)$  and let  $0 \neq h(\mathbf{a}) \in H$ . We have an equality of type  $h(\mathbf{a}) = a(\mathbf{a}, \mathbf{x})f(\mathbf{a}, \mathbf{x}) + b(\mathbf{a}, \mathbf{x})D(\mathbf{a}, \mathbf{x})$  with  $f(\mathbf{a}, \mathbf{x}) \in I(\mathbf{a}, \mathbf{x})$ , and hence an equality  $1 = \frac{a(\mathbf{a}, \mathbf{x})}{h(\mathbf{a})}f(\mathbf{a}, \mathbf{x}) + \frac{b(\mathbf{a}, \mathbf{x})}{h(\mathbf{a})}D(\mathbf{a}, \mathbf{x})$  in  $J(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}]$ . For every  $\alpha \in \mathcal{S}$  such that  $h(\alpha) \neq 0$  the equality implies that the corresponding complete intersection has no common zeros with the determinant of its Jacobian matrix, hence it is

smooth. Conversely, assume that  $H = (0)$ . Then the canonical  $K$ -algebra homomorphism  $K[\mathbf{a}] \rightarrow K[\mathbf{a}, \mathbf{x}]/J(\mathbf{a}, \mathbf{x})$  is injective and hence it induces a morphism  $\text{Spec}(K[\mathbf{a}, \mathbf{x}]/J(\mathbf{a}, \mathbf{x})) \rightarrow \mathbb{A}_K^m$  of affine schemes which is dominant. It means that for a generic point of  $\mathbb{A}_K^m$ , the scheme  $\text{Spec}(K[\mathbf{a}, \mathbf{x}]/J(\mathbf{a}, \mathbf{x}))$  is not empty, and hence the corresponding complete intersection  $\text{Spec}(K[\mathbf{a}, \mathbf{x}]/I(\mathbf{a}, \mathbf{x}))$  is not smooth.  $\square$

The following example illustrates these results.

**Example 2.13.** Let us consider the polynomials  $f_1 = x_1^2 + x_2^2 - 1$ ,  $f_2 = x_2^2 + x_1$  in  $\mathbb{C}[x_1, x_2]$  and the ideal  $I = (f_1, f_2)$  generated by them. It is a zero-dimensional complete intersection and we embed it into  $I(\mathbf{a}, \mathbf{x}) = (x_1^2 + a_1x_2^2 - 1, x_2^2 + a_2x_1)$ . It is a free family over  $\mathbb{A}_{\mathbb{C}}^2$ , and the multiplicity of each fiber is 4. We compute  $D(\mathbf{a}, \mathbf{x}) = \det(\text{Jac}_F(\mathbf{a}, \mathbf{x}))$  and get  $D(\mathbf{a}, \mathbf{x}) = -2a_1a_2x_2 + 4x_1x_2$ . We let

$$J(\mathbf{a}, \mathbf{x}) = I(\mathbf{a}, \mathbf{x}) + (D(\mathbf{a}, \mathbf{x})) = (x_1^2 + a_1x_2^2 - 1, x_2^2 + a_2x_1, -2a_1a_2x_2 + 4x_1x_2)$$

A computation with CoCoA of  $\text{Elim}([x_1, x_2], J)$  yields  $(\frac{1}{2}a_1^2a_2^3 + 2a_2)$ , and hence  $J(\mathbf{a}, \mathbf{x}) \cap K[\mathbf{a}] = (\frac{1}{2}a_1^2a_2^3 + 2a_2)$ . According to the theorem, if  $\mathcal{U}$  is the complement in  $\mathbb{A}_{\mathbb{C}}^2$  of the curve defined by  $\frac{1}{2}a_1^2a_2^3 + 2a_2 = 0$ , then  $\mathcal{U}$  is an  $I$ -smooth subscheme of  $\mathbb{A}_{\mathbb{C}}^2$ . On the other hand, the curve has three components,  $a_2 = 0$ , and  $a_1a_2 \pm 2i = 0$ . If  $a_2 = 0$  then the corresponding ideal is  $(x_1^2 - 1, x_2^2)$  which is not smooth. If  $a_1a_2 \pm 2i = 0$ , then the corresponding ideals are  $(x_1^2 \mp \frac{2i}{a_2}x_2^2 - 1, x_2^2 + a_2x_1)$  which can be written as  $((x_1 \pm i)^2, x_2^2 + a_2x_1)$  and hence are not smooth.

Let us now consider the zero-dimensional complete intersection described by the ideal  $I = (f_1, f_2)$  where  $f_1 = x_1^2 + x_2^2$ ,  $f_2 = x_2^2 + x_1$ . We embed it into the family  $I(\mathbf{a}, \mathbf{x}) = (x_1^2 - a_1x_2^2, x_2^2 + a_2x_1)$ . As before, it is a free family over  $\mathbb{A}_{\mathbb{C}}^2$ , and the multiplicity of each fiber is 4. We compute  $D(\mathbf{a}, \mathbf{x}) = \det(\text{Jac}_F(\mathbf{a}, \mathbf{x}))$  and get  $D(\mathbf{a}, \mathbf{x}) = 2a_1a_2x_2 + 4x_1x_2$ . The computation of  $\text{Elim}([x, y], J)$  yields  $(0)$ , and hence there is no subscheme of  $\mathbb{A}_K^2$  which is  $I$ -smooth. Indeed, for  $a_2 \neq 0$  we have  $I(\mathbf{a}, \mathbf{x}) = (x_1 + \frac{1}{a_2}x_2^2, \frac{1}{a_2^2}x_2^4 - a_1x_2^2)$  which is not smooth. Incidentally, we observe that also for  $a_2 = 0$  the corresponding zero-dimensional complete intersection is not smooth.

The following example illustrates other subtleties related to the theorem.

**Example 2.14. (Example 2.8 continued)**

We consider the family  $I(\mathbf{a}, \mathbf{x}) = (ax^3 - y, f_2)$  for  $a \neq 0$  of Example 2.8, compute  $D(\mathbf{a}, \mathbf{x}) = \det(\text{Jac}_F(\mathbf{a}, \mathbf{x}))$  and get  $D(\mathbf{a}, \mathbf{x}) = 9ax^3y^2 + 1512ax^4 - 1098ax^3y + 126ax^2y^2 + 1950ax^3 - 882ax^2y + y^3 + 399ax^2 + 1008xy - 183y^2 - 1008x + 650y - 468$ . We let  $J(\mathbf{a}, \mathbf{x}) = I(\mathbf{a}, \mathbf{x}) + (D(\mathbf{a}, \mathbf{x}))$  and get  $J(\mathbf{a}, \mathbf{x}) \cap K[\mathbf{a}] = (h(\mathbf{a}))$  where

$$\begin{aligned} h(\mathbf{a}) = & a^9 - \frac{738170716516748}{7749152384519}a^8 + \frac{218039463835944563500746}{91409877182005574647}a^7 - \frac{166557011563009981474061668}{31353587873427912103921}a^6 \\ & - \frac{276169260891419750846552207}{31353587873427912103921}a^5 + \frac{986809115998719019081678896}{31353587873427912103921}a^4 - \frac{63247607413926237871517952}{31353587873427912103921}a^3 \\ & - \frac{1316764479863922379654192128}{31353587873427912103921}a^2 + \frac{317872550804296477704192}{13058553883143653521}a - \frac{974975584016793600000}{266501099655992929} \end{aligned}$$

Therefore, if  $\mathcal{U}$  denotes the complement in  $\mathbb{A}_K^1$  of the zeros of  $h(a)$ , the theorem says that it is a Zariski-open  $I$ -smooth subscheme. However, we have already seen in Example 2.8 that  $a = 0$  (the origin is in  $\mathcal{U}$ ) is not in the free locus: we observe that the corresponding complete intersection is smooth, but it has only two points. The other subtlety is that the Bézout number of the family is  $3 \times 4 = 12$ , but if we substitute  $y = ax^3$  into  $f_2$  we get a univariate polynomial of degree 10. The two *missing* points are at infinity. No member of the family

represents twelve points. The final remark is that if we move the parameter  $a$  in the locus described by  $ah(a) \neq 0$  we always get a smooth complete intersection of 10 points. If  $K = \mathbb{C}$  the ten points have complex coordinates, some of them are real, but there are no values of  $a$  for which all the 10 points are real. The reason is that if  $r_1 = \frac{-1+\sqrt{3}i}{2}$ ,  $r_2 = \frac{-1-\sqrt{3}i}{2}$  are the two complex roots of  $x^2 + x + 1 = 0$ , then two of the ten points are  $(r_1, r_1^3)$ ,  $(r_2, r_2^3)$  which are not real points (see Theorem 2.20 and Example 2.22).

Combining Theorem 2.12 and Proposition 2.6 we get a method to select a Zariski-open subscheme of the parameter space over which all the fibers are smooth complete intersections of constant multiplicity (see [18] for similar results). Before describing the algorithm, we need a definition which captures this concept.

**Definition 2.15.** With the above notation, a dense Zariski-open subscheme  $\mathcal{U}$  of  $\mathcal{S}$  such  $\Phi^{-1}(\mathcal{U}) \rightarrow \mathcal{U}$  is smooth and free is said to be an *I-optimal* subscheme of  $\mathcal{S}$ .

**Corollary 2.16.** Let  $\mathcal{S} = \mathbb{A}_K^m$  and consider the following sequence of instructions.

- (1) Compute  $D(\mathbf{a}, \mathbf{x}) = \det(\text{Jac}_F(\mathbf{a}, \mathbf{x}))$ .
- (2) Let  $J(\mathbf{a}, \mathbf{x}) = I(\mathbf{a}, \mathbf{x}) + (D(\mathbf{a}, \mathbf{x}))$  and compute  $H = J(\mathbf{a}, \mathbf{x}) \cap K[\mathbf{a}]$ .
- (3) If  $H = (0)$  return “There is no  $I$ -smooth subscheme of  $\mathbb{A}_K^m$ ” and stop.
- (4) Choose  $h(\mathbf{a}) \in H \setminus 0$  and let  $\mathcal{U}_1 = \mathbb{A}_K^m \setminus \{\alpha \in \mathbb{A}_K^m \mid h(\alpha) = 0\}$ .
- (5) Choose a term ordering  $\sigma$  on  $\mathbb{T}^n$  and compute the reduced  $\sigma$ -Gröbner basis  $G(\mathbf{a}, \mathbf{x})$  of  $I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}]$ .
- (6) Let  $T = \mathbb{T}^n \setminus \text{LT}_\sigma(I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}])$ , compute the cardinality of  $T$  and call it  $\mu$ ; then compute the least common multiple of all the denominators of the coefficients of the polynomials in  $G(\mathbf{a}, \mathbf{x})$ , and call it  $d(\mathbf{a})$ ; finally, let  $\mathcal{U}_2 = \mathbb{A}_K^m \setminus \{\alpha \in \mathbb{A}_K^m \mid d(\alpha) \neq 0\}$  and let  $\mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2$ .
- (7) Return  $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}, T, \mu$ .

This is an algorithm which returns  $\mathcal{U}_1$  which is  $I$ -smooth,  $\mathcal{U}_2$  which is  $I$ -free,  $\mathcal{U}$  which is  $I$ -optimal,  $T$  which provides a basis as  $K$ -vector spaces of all the fibers over  $\mathcal{U}_2$ , and  $\mu$  which is the multiplicity of all the fibers over  $\mathcal{U}_2$ .

*Proof.* It suffices to combine Theorem 2.12 and Proposition 2.6.  $\square$

**Example 2.17.** We consider the ideal  $I = (f_1, f_2)$  of  $K[x, y]$  where  $f_1 = xy - 6$ ,  $f_2 = x^2 + y^2 - 13$ . It is a zero-dimensional complete intersection and we embed it into the family  $I(\mathbf{a}, \mathbf{x}) = (a_1xy + a_2, a_3x^2 + a_4y^2 + a_5)$ . We compute the reduced DegRevLex-Gröbner basis of  $I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[\mathbf{x}]$  and get

$$\left\{x^2 + \frac{a_4}{a_3}y^2 + \frac{a_5}{a_3}, \quad xy + \frac{a_2}{a_1}, \quad y^3 - \frac{a_2a_3}{a_1a_4}x + \frac{a_1a_5}{a_1a_4}y\right\}$$

according to the above results, a free locus is given by  $a_1a_3a_4 \neq 0$ . Now we compute  $D(\mathbf{a}, \mathbf{x}) = \det(\text{Jac}_F(\mathbf{a}, \mathbf{x}))$  and get  $D(\mathbf{a}, \mathbf{x}) = -2a_1a_3x^2 + 2a_1a_4y^2$ .

We let  $J(\mathbf{a}, \mathbf{x}) = I(\mathbf{a}, \mathbf{x}) + (D(\mathbf{a}, \mathbf{x}))$  and compute  $J(\mathbf{a}, \mathbf{x}) \cap K[\mathbf{a}]$ . We get the principal ideal generated by  $a_2^2a_3a_4 - \frac{1}{4}a_1^2a_5^2$ . In conclusion, an  $I$ -optimal subscheme is  $\mathcal{U} = \mathbb{A}_K^5 \setminus F$  where  $F$  is the closed subscheme defined by the equation  $a_1a_3a_4(a_2^2a_3a_4 - \frac{1}{4}a_1^2a_5^2) = 0$ , and  $\mu = 4$ .

**Definition 2.18.** We say that a **point is complex** if its coordinates are complex numbers, and we say that a **point is real** if its coordinates are real numbers.

The following example illustrates the fact that even if we start with a set of real points, a zero-dimensional complete intersection which contains them may also contain complex non-real points.

**Example 2.19.** Let  $\mathbb{X}$  be the set of the 10 real points  $\{(-1, -1), (2, 8), (-2, -8), (3, 27), (-3, -27), (4, 64), (5, 125), (-5, -125), (6, 216), (-6, -216)\}$ . A zero-dimensional complete intersection containing  $\mathbb{X}$  is  $\{f_1, f_2\}$  where  $f_1 = y - x^3$  and  $f_2 = x^2y^2 - 1/4095y^4 + 1729/15x^2y - 74/15xy^2 + 1/15y^3 - 8832/5x^2 + 5852/15xy - 10754/315y^2 + 2160x - 4632/5y + 250560/91$ . Let  $I$  denote the vanishing ideal of the 10 points and let  $J$  denote the ideal generated by  $\{f_1, f_2\}$ . The colon ideal  $J : I$  defines the residual intersection. Since  $J$  is the intersection of a cubic and a quartic curve, the residual intersection is a zero-dimensional scheme of multiplicity 2. Indeed, a computation (performed with CoCoA) shows that  $J : I$  is generated by  $(x + 1/78y - 87/26, y^2 - 756y + 658503)$ . Since  $756^2 - 4 \cdot 658503 = -2062476 < 0$ , the two extra points on the zero-dimensional complete intersection are complex, non real points.

**Theorem 2.20.** Let  $\mathbf{f}(\mathbf{x})$  be a zero-dimensional complete intersection in  $\mathbb{R}[\mathbf{x}]$  and let  $\mathbf{f}(\mathbf{a}, \mathbf{x}) \in \mathbb{R}[\mathbf{a}, \mathbf{x}]$  be a zero-dimensional family containing  $\mathbf{f}(\mathbf{x})$ . Let  $I$  be the ideal in  $\mathbb{R}[\mathbf{x}]$  generated by  $\mathbf{f}(\mathbf{x})$ , assume that there exists an  $I$ -optimal subscheme  $\mathcal{U}$  of  $\mathbb{A}_{\mathbb{R}}^m$ , and let  $\alpha_I \in \mathcal{U}$  be the point in the parameter space which corresponds to  $I$ . If  $\mu_{\mathbb{R}, I}$  is the number of distinct real points in the fiber over  $\alpha_I$  (i.e. zeroes of  $I$ ), then there exist an open semi-algebraic subscheme  $\mathcal{V}$  of  $\mathcal{U}$  such that for every  $\alpha \in \mathcal{V}$  the number of real points in the fiber over  $\alpha$  is  $\mu_{\mathbb{R}, I}$ .

*Proof.* We consider the ideal  $\mathcal{I} = I(\mathbf{a}, \mathbf{x})\mathbb{R}(\mathbf{a})[\mathbf{x}]$ . It is zero-dimensional and the field  $\mathbb{R}(\mathbf{a})$  is infinite. Since a linear change of coordinates does not change the problem, we may assume that  $\mathcal{I}$  is in  $x_n$ -normal position (see [15], Section 3.7). Moreover, we have already observed (see Remark 2.7) that in Proposition 2.6 the choice of  $\sigma$  is arbitrary. We choose  $\sigma = \text{Lex}$  and hence the reduced  $\text{Lex}$ -Gröbner basis of  $\mathcal{I}$  has the shape prescribed by the Shape Lemma (see [15] Theorem 3.7.25). Therefore there exists a univariate polynomial  $h_{\mathbf{a}} \in \mathbb{R}(\mathbf{a})[x_n]$  whose degree is the multiplicity of both the generic fiber and the fiber over  $\alpha_I$ , which is the number of complex zeros of  $I$ . Due to the shape of the reduced Gröbner basis, a point is real if and only if its  $x_n$ -coordinate is real. Therefore it suffices to prove the following statement: given a univariate square-free polynomial  $h_{\mathbf{a}} \in \mathbb{R}(\mathbf{a})[x_n]$  such that  $h_{\alpha_I}$  has exactly  $\mu_{\mathbb{R}, I}$  real roots, there exists an open semi-algebraic subset of  $A_{\mathbb{R}}^m$  such that for every point  $\alpha$  in it, the polynomial  $h_{\alpha}$  has exactly  $\mu_{\mathbb{R}, I}$  real roots. This statement follows from [5], Theorem 5.12 where it is shown that for every root there exists an open semi-algebraic set in  $A_{\mathbb{R}}^m$  which isolates the root. Since complex non-real roots have to occur in conjugate pairs, this implies that real roots stay real.  $\square$

Let us see some examples.

**Example 2.21.** We consider the ideal  $I = (xy - 2y^2 + 2y, x^2 - y^2 - 2x)$  in  $\mathbb{R}[x, y]$ , and we embed it into the family  $I(\mathbf{a}, \mathbf{x}) = (xy - ay^2 + ay, x^2 - y^2 - 2x)$ . We compute the reduced  $\text{Lex}$ -Gröbner basis of  $I(\mathbf{a}, \mathbf{x})\mathbb{R}(\mathbf{a})[\mathbf{x}]$  and get

$$\{x^2 - 2x - y^2, xy - ay^2 + ay, y^3 - \frac{2a}{a-1}y^2 + \frac{a^2+2a}{a^2-1}y\}$$



Applying the algorithm illustrated in Corollary 2.16 we get an  $I$ -smooth subscheme of  $\mathbb{A}_{\mathbb{R}}^1$  for  $a(a+2) \neq 0$ , and an  $I$ -free subscheme for  $(a-1)(a+1) \neq 0$ . For  $a$  different from 0,  $-2$ ,  $1$ ,  $-1$  we have an  $I$ -optimal subscheme and the multiplicity is 4.

Our ideal  $I$  is obtained for  $a = 2$ , and hence it lies over the optimal subscheme. It has multiplicity 4 and the four zeros are real.

The computed Lex-Gröbner basis does not have the shape prescribed by the Shape Lemma, so we perform a linear change of coordinates by setting  $x = x + y$ ,  $y = x - y$ . We compute the reduced Lex-Gröbner basis and get

$$\{x + 4\frac{a+1}{a-1}y^3 - 2\frac{a+1}{a-1}y^2 - \frac{3a+1}{a-1}y, \quad y^4 - y^3 - \frac{1}{2}\frac{a}{a+1}y^2 + \frac{1}{2}\frac{a}{a+1}y\}$$

It has the good shape, so we can use the polynomial

$$h_{\mathbf{a}} = y^4 - y^3 - \frac{1}{2}\frac{a}{a+1}y^2 + \frac{1}{2}\frac{a}{a+1}y = y(y-1)(y^2 - \frac{1}{2}\frac{a}{a+1})$$

We get the following result.

- For  $a < -1$ ,  $a \neq -2$  there are 4 real points.
- For  $-1 < a < 0$  there are 2 real points.
- For  $a > 0$ ,  $a \neq 1$  there are 4 real points.

To complete our analysis, let us see what happens at the *bad* points 0,  $-2$ ,  $1$ ,  $-1$ .

At 0 the primary decomposition of the ideal  $I_0$  is  $(x-2, y) \cap (y^2 + 2x, xy, x^2)$ , hence the fiber consists in the simple point  $(2, 0)$  and a triple point at  $(0, 0)$ .

At  $-2$  we see that  $(x + \frac{2}{3}, y - \frac{4}{3}) \cap (x, y) \cap (x-2, y^2)$  is the primary decomposition of the ideal  $I_{-2}$ , and hence the fiber consists in the simple point  $(-\frac{2}{3}, \frac{4}{3})$ , the simple point  $(0, 0)$  and a double point at  $(2, 0)$ .

At  $-1$  the primary decomposition of the ideal  $I_{-1}$  is  $(x, y) \cap (x-2, y)$ , hence the fiber consists of the two simple real points  $(0, 0)$  and  $(2, 0)$ .

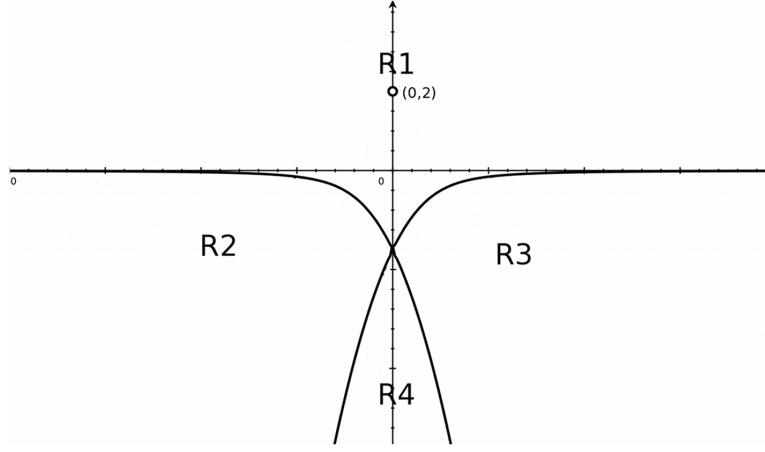
At 1 we see that  $(x, y) \cap (x-2, y) \cap (x + \frac{1}{4}, y - \frac{3}{4})$  is the primary decomposition of the ideal  $I_1$ , hence the fiber consists of the three simple real points  $(0, 0)$ ,  $(2, 0)$ ,  $(-\frac{1}{4}, \frac{3}{4})$ .

**Example 2.22.** We consider the ideal  $I = (xy + 1, x^2 + y^2 - 5)$  in  $\mathbb{R}[x, y]$ , and we embed it into the family  $I(\mathbf{a}, x, y) = (xy + a_1x + 1, x^2 + y^2 + a_2)$ . We compute the reduced Lex-Gröbner basis of  $I(\mathbf{a}, \mathbf{x})K(\mathbf{a})[x, y]$  and get  $G(\mathbf{a}, x, y) = \{g_1, g_2\}$  where

$$\begin{aligned} g_1 &= x - y^3 - a_1y^2 - a_2y - a_1a_2, \\ g_2 &= y^4 + 2a_1y^3 + (a_1^2 + a_2)y^2 + 2a_1a_2y + (a_1^2a_2 + 1) \end{aligned}$$

which has the shape prescribed by the Shape Lemma (see [15] Theorem 3.7.25). There is no condition for the free locus, and  $D(\mathbf{a}, x, y) = \det(\text{Jac}_F(\mathbf{a}, x, y)) = -2x^2 + 2y^2 + 2a_1y$ . We let  $J(\mathbf{a}, x, y) = I(\mathbf{a}, x, y) + (D(\mathbf{a}, x, y))$  and compute  $J(\mathbf{a}, x, y) \cap K[\mathbf{a}]$ . We get the principal ideal generated by the following polynomial  $h(\mathbf{a}) = a_1^6a_2 + 3a_1^4a_2^2 + a_1^4 + 3a_1^2a_2^3 + 20a_1^2a_2 + a_2^4 - 8a_2^2 + 16$ . An  $I$ -optimal subscheme is  $\mathcal{U} = \mathbb{A}_{\mathbb{R}}^4 \setminus F$  where  $F$  is the closed subscheme defined by the equation  $h(\mathbf{a}) = 0$ , and we observe that  $\mu = 4$ .

At this point we know that for  $h(\mathbf{a}) \neq 0$  each fiber is smooth and has multiplicity 4, hence it consists of 4 distinct complex points. What about real points?



The real curve defined by  $h(\mathbf{a}) = 0$  is shown in the above picture. It is the union of two branches and the isolated point  $(0, 2)$ . The upper region R1 (with the exception of the point  $(0, 2)$ ) corresponds to the ideals in the family whose zeros are four complex non-real points. The regions R2 and R3 correspond to the ideals whose zeros are two complex non-real points and two real points. The region R4 corresponds to the ideals whose zeros are four real points. To describe the four regions algebraically, we use the Sturm-Habicht sequence (see [12]) of  $g_2 \in \mathbb{R}(\mathbf{a})[y]$ . The leading monomials are  $y^4, 4y^3, 4r(\mathbf{a})y^2, -8\ell(\mathbf{a})y, 16h(\mathbf{a})$  where  $r(\mathbf{a}) = a_1^2 - 2a_2$ ,  $\ell(\mathbf{a}) = a_1^4 a_2 + 2a_1^2 a_2^2 + 2a_1^2 + a_2^3 - 4a_2$ . To get the total number of real roots we count the sign changes in the sequence at  $-\infty$  and  $+\infty$ ; in particular, we observe that in the parameter space the ideal  $I$  corresponds to the point  $(0, -5)$  which belongs to the region R4. We get

$$R4 = \{\alpha \in \mathbb{R}^2 \mid r(\alpha) > 0, \ell(\alpha) < 0, h(\alpha) > 0\}$$

which is semi-algebraic open, not Zariski-open.

### 3. CONDITION NUMBERS

In this section we introduce a notion of *condition number* for zero-dimensional smooth complete intersections in  $\mathbb{R}[\mathbf{x}]$ ; the aim is to give a measure of the sensitivity of its real roots with respect to small perturbations of the input data, that is small changes of the coefficients of the involved polynomials.

The section starts with the recall of well-known facts about numerical linear algebra. We let  $m, n$  be positive integers and let  $\text{Mat}_{m \times n}(\mathbb{R})$  be the set of  $m \times n$  matrices with entries in  $\mathbb{R}$ ; if  $m = n$  we simply write  $\text{Mat}_n(\mathbb{R})$ .

**Definition 3.1.** Let  $M = (m_{ij})$  be a matrix in  $\text{Mat}_{m \times n}(\mathbb{R})$ ,  $v = (v_1, \dots, v_n)$  a vector in  $\mathbb{R}^n$  and  $\|\cdot\|$  a vector norm.

- (a) Let  $r \geq 1$  be a real number; the  **$r$ -norm** on the vector space  $\mathbb{R}^n$  is defined by the formula  $\|v\|_r = (\sum_{i=1}^n |v_i|^r)^{\frac{1}{r}}$  for every  $v \in \mathbb{R}^n$ .
- (b) The **infinity norm** on  $\mathbb{R}^n$  is defined by the formula  $\|v\|_\infty = \max_i |v_i|$ .
- (c) The **spectral radius**  $\varrho(M)$  of the matrix  $M$  is defined by the formula  $\varrho(M) = \max_i |\lambda_i|$ , where the  $\lambda_i$  are the *complex* eigenvalues of  $M$ .

- (d) The real function defined on  $\text{Mat}_{m \times n}(\mathbb{R})$  by  $M \mapsto \max_{\|v\|=1} \|Mv\|$  is a matrix norm called the **matrix norm induced** by  $\|\cdot\|$ . A matrix norm induced by a vector norm is called an **induced matrix norm**.
- (e) The matrix norm induced by  $\|\cdot\|_1$  is given by the following formula  $\|M\|_1 = \max_j (\sum_i |m_{ij}|)$ . The matrix norm induced by  $\|\cdot\|_\infty$  is given by the formula  $\|M\|_\infty = \max_i (\sum_j |m_{ij}|)$ . Finally, the matrix norm induced by  $\|\cdot\|_2$  is given by the formula  $\|M\|_2 = \max_i (\sigma_i)$  where the  $\sigma_i$  are singular values of  $M$ .

If no confusion arises, from now on we will use the symbol  $\|\cdot\|$  to denote both a vector norm and a matrix norm. We recall some facts about matrix norms (see for instance [4], [13]).

**Proposition 3.2.** *Let  $M$  be a matrix in  $\text{Mat}_n(\mathbb{R})$ , let  $I$  be the identity matrix of type  $n$  and let  $\|\cdot\|$  be an induced matrix norm on  $\text{Mat}_n(\mathbb{R})$ . If the matrix  $I + M$  is invertible then  $(1 - \|M\|) \|(I + M)^{-1}\| \leq 1$ .*

**Proposition 3.3.** *Let  $M \in \text{Mat}_{m \times n}(\mathbb{R})$  and denote by  $M_i$  the  $i$ -th row of  $M$ . Let  $r_1 \geq 1, r_2 \geq 1$  be real numbers such that  $\frac{1}{r_1} + \frac{1}{r_2} = 1$ ; then*

$$\max_i \|M_i\|_{r_2} \leq \|M\|_{r_1} \leq m^{1/r_1} \max_i \|M_i\|_{r_2}$$

*In particular, for  $r_1 = r_2 = 2$*

$$\max_i \|M_i\|_2 \leq \|M\|_2 \leq \sqrt{m} \max_i \|M_i\|_2$$

This introductory part ends with the recollection of some facts about the polynomial ring  $K[\mathbf{x}]$ . In particular, given  $\eta = (\eta_1, \dots, \eta_n) \in \mathbb{N}^n$  we denote by  $|\eta|$  the number  $\eta_1 + \dots + \eta_n$ , by  $\eta!$  the number  $\eta_1! \dots \eta_n!$ , and by  $\mathbf{x}^\eta$  the power product  $x_1^{\eta_1} \dots x_n^{\eta_n}$ .

**Definition 3.4.** Let  $p$  be a point of  $K^n$ ; the  $K$ -linear map on  $K[\mathbf{x}]$  defined by  $f \mapsto f(p)$  is called the **evaluation map** associated to  $p$  and denoted by  $\text{ev}_p(f)$ .

**Definition 3.5.** Let  $d$  be a nonnegative integer, let  $r \geq 1$  be a real number, let  $p$  be a point of  $\mathbb{R}^n$  and let  $g(\mathbf{x})$  be a polynomial in  $\mathbb{R}[\mathbf{x}]$ .

- (a) The formal Taylor expansion of  $g(\mathbf{x})$  at  $p$  is given by the following expression:  $g(\mathbf{x}) = \sum_{|\eta| \geq 0} \frac{1}{\eta!} \frac{\partial^\eta g}{\partial \mathbf{x}^\eta}(p)(\mathbf{x} - p)^\eta$ .
- (b) The polynomial  $\sum_{|\eta| \geq d} \frac{1}{\eta!} \frac{\partial^\eta g}{\partial \mathbf{x}^\eta}(p)(\mathbf{x} - p)^\eta$  is denoted by  $g^{\geq d}(\mathbf{x}, p)$ .
- (c) The  **$r$ -norm of  $g(\mathbf{x})$  at  $p$**  is defined as the  $r$ -norm of the vector  $\frac{\partial g}{\partial \mathbf{x}}(p)$ . If  $\|\frac{\partial g}{\partial \mathbf{x}}(p)\|_r = 1$  then  $g(\mathbf{x})$  is called **unitary at  $p$** .

We use the following formulation of Taylor's theorem.

**Proposition 3.6.** *Let  $p$  be a point of  $\mathbb{R}^n$  and let  $g(\mathbf{x})$  be a polynomial in  $\mathbb{R}[\mathbf{x}]$ . For every point  $q \in \mathbb{R}^n$  we have*

$$g(q) = g(p) + \text{Jac}_g(p)(q - p) + \frac{1}{2}(q - p)^t H_g(\xi)(q - p)$$

*where  $\xi$  is a point of the line connecting  $p$  to  $q$  and  $H_g(\xi)$  is the Hessian matrix of  $g$  at  $\xi$ .*

Given  $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$ , a zero-dimensional smooth complete intersection in  $\mathbb{R}[\mathbf{x}]$ , we introduce a notion of admissible perturbation of  $\mathbf{f}(\mathbf{x})$ . Roughly speaking, the polynomial set  $\varepsilon(\mathbf{x}) = \{\varepsilon_1(\mathbf{x}), \dots, \varepsilon_n(\mathbf{x})\} \subset \mathbb{R}[\mathbf{x}]$  is considered to

be an admissible perturbation of  $\mathbf{f}(\mathbf{x})$  if the real solutions of  $(\mathbf{f} + \varepsilon)(\mathbf{x}) = 0$  are nonsingular and derive from perturbations of the real solutions of  $\mathbf{f}(\mathbf{x}) = 0$ . Using the results of Section 2 we formalize this concept as follows.

**Definition 3.7.** Let  $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$  be a zero-dimensional smooth complete intersection in  $\mathbb{R}[\mathbf{x}]$ , let  $\mu_{\mathbb{R}, I}$  be the number of real solutions of  $\mathbf{f}(\mathbf{x}) = 0$ , and let  $\varepsilon(\mathbf{x}) = \{\varepsilon_1(\mathbf{x}), \dots, \varepsilon_n(\mathbf{x})\}$  be a set of polynomials in  $\mathbb{R}[\mathbf{x}]$ . Suppose that the assumptions of Theorem 2.20 are satisfied, let  $\mathcal{V} \subset \mathbb{A}_{\mathbb{R}}^m$  be an open semi-algebraic subset of  $\mathcal{U}$  such that  $\alpha_I \in \mathcal{V}$ , and for every  $\alpha \in \mathcal{V}$  the number of real roots of  $\mathbf{f}(\alpha, \mathbf{x}) = 0$  is equal to  $\mu_{\mathbb{R}, I}$ . If there exists  $\alpha \in \mathcal{V}$  such that  $(\mathbf{f} + \varepsilon)(\mathbf{x}) = \mathbf{f}(\alpha, \mathbf{x})$ , then  $\varepsilon(\mathbf{x})$  is called an **admissible perturbation** of  $\mathbf{f}(\mathbf{x})$ .

Henceforth we let  $\varepsilon(\mathbf{x}) = \{\varepsilon_1(\mathbf{x}), \dots, \varepsilon_n(\mathbf{x})\}$  be an admissible perturbation of  $\mathbf{f}(\mathbf{x})$ , and let  $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}) = \{p_1, \dots, p_{\mu_{\mathbb{R}, I}}\}$ ,  $\mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \varepsilon) = \{r_1, \dots, r_{\mu_{\mathbb{R}, I}}\}$  be the sets of real solutions of  $\mathbf{f}(\mathbf{x}) = 0$  and  $(\mathbf{f} + \varepsilon)(\mathbf{x}) = 0$  respectively. We consider each  $r_i$  as a perturbation of the root  $p_i$ , hence we write  $r_i = p_i + \Delta p_i$  for  $i = 1, \dots, \mu_{\mathbb{R}, I}$ .

Now we concentrate on a single element  $p$  of  $\mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ .

**Corollary 3.8.** *Let  $p$  be one of the real solutions of  $\mathbf{f} = 0$ , and  $p + \Delta p$  the corresponding real solution of  $\mathbf{f} + \varepsilon = 0$ . Then we have*

$$(1) \quad 0 = (\mathbf{f} + \varepsilon)(p + \Delta p) = \varepsilon(p) + \text{Jac}_{\mathbf{f} + \varepsilon}(p)\Delta p + (v_1(\xi_1), \dots, v_n(\xi_n))^t$$

where  $\xi_1, \dots, \xi_n$  are points on the line which connects the points  $p$  and  $p + \Delta p$ , and  $v_j(\xi_j) = \frac{1}{2}\Delta p^t H_{f_j + \varepsilon_j}(\xi_j)\Delta p$  for each  $j = 1, \dots, n$ .

*Proof.* It suffices to put  $q = p + \Delta p$ , apply the formula of Proposition 3.6 to the polynomial system  $(\mathbf{f} + \varepsilon)(\mathbf{x})$ , and use the fact that  $\mathbf{f}(p) = 0$ .  $\square$

**Example 3.9.** We consider the zero-dimensional smooth complete intersection  $\mathbf{f} = \{f_1, f_2\}$  where  $f_1 = xy - 6$ ,  $f_2 = x^2 + y^2 - 13$  and observe that  $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}) = \{(-3, -2), (3, 2), (-2, -3), (2, 3)\}$ . The set  $\mathbf{f}(\mathbf{x})$  is embedded into the following family  $F(\mathbf{a}, \mathbf{x}) = \{xy + a_1, x^2 + a_2y^2 + a_3\}$ .

The semi-algebraic open set

$$\mathcal{V} = \{\alpha \in \mathbb{R}^3 \mid \alpha_3^2 - 4\alpha_1^2\alpha_2 > 0, \alpha_2 > 0, \alpha_3 < 0\}$$

is a subset of the  $I$ -optimal scheme  $\mathcal{U} = \{\alpha \in A_{\mathbb{R}}^3 \mid \alpha_2(\alpha_3^2 - 4\alpha_1^2\alpha_2) \neq 0\}$ . Moreover, it contains the point  $\alpha_I = (-6, 1, -13)$ , and the fiber over each  $\alpha \in \mathcal{V}$  consists of 4 real points. The set  $\varepsilon(\mathbf{x}) = \{\delta_1, \delta_2y^2 + \delta_3\}$ , with  $\delta_i \in \mathbb{R}$ , is an admissible perturbation of  $\mathbf{f}(\mathbf{x})$  if and only if the conditions  $(\delta_3 - 13)^2 - 4(\delta_1 - 6)^2(\delta_2 + 1) > 0$ ,  $\delta_2 > -1$ , and  $\delta_3 < 13$  are satisfied. Since the values  $\delta_1 = 2$ ,  $\delta_2 = \frac{5}{4}$ , and  $\delta_3 = 0$  satisfy the previous conditions, the polynomial set  $\varepsilon(\mathbf{x}) = \{2, \frac{5}{4}y^2\}$  is an admissible perturbation of  $\mathbf{f}(\mathbf{x})$ . The real roots of  $(\mathbf{f} + \varepsilon)(\mathbf{x}) = 0$  are

$$\mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \varepsilon) = \left\{ \left(-3, -\frac{4}{3}\right), \left(3, \frac{4}{3}\right), (-2, -2), (2, 2) \right\}$$

For each  $r_i \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \varepsilon)$  the matrix  $\text{Jac}_{\mathbf{f} + \varepsilon}(r_i)$  is invertible, as predicted by the theory. On the contrary, by evaluating  $\text{Jac}_{\mathbf{f} + \varepsilon}(\mathbf{x})$  at the third and the fourth point of  $\mathcal{Z}_{\mathbb{R}}(\mathbf{f})$  we obtain a singular matrix. This is an obstruction to the development of the theory which suggests further restrictions (see the following discussion).

Our idea is to evaluate  $\Delta p$  using equation (1) of Corollary 3.8. However, while the assumption that  $\varepsilon(\mathbf{x})$  is an admissible perturbation of  $\mathbf{f}(\mathbf{x})$  combined with the Jacobian criterion guarantee the non singularity of the matrix  $\text{Jac}_{\mathbf{f} + \varepsilon}(p + \Delta p)$ ,

they do not imply the non singularity of the matrix  $\text{Jac}_{\mathbf{f}+\varepsilon}(p)$ , as we have just seen in Example 3.9. The next step is to find a criterion which guarantees the non singularity of  $\text{Jac}_{\mathbf{f}+\varepsilon}(p)$ .

**Lemma 3.10.** *If  $\|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\| < 1$  then  $\text{Jac}_{\mathbf{f}+\varepsilon}(p)$  is invertible.*

*Proof.* By assumption  $p$  is a nonsingular root of  $\mathbf{f}(\mathbf{x}) = 0$ , hence  $\text{Jac}_{\mathbf{f}}(p)$  is invertible and so  $\text{Jac}_{\mathbf{f}+\varepsilon}(p)$  can be rewritten as  $\text{Jac}_{\mathbf{f}+\varepsilon}(p) = \text{Jac}_{\mathbf{f}}(p) + \text{Jac}_{\varepsilon}(p) = \text{Jac}_{\mathbf{f}}(p) (I + \text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p))$ . Consequently, it suffices to show that the matrix  $I + \text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)$  is invertible. And we achieve it by proving that the spectral radius  $\varrho(\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p))$  is smaller than 1. We have  $\varrho(\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)) \leq \|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\| < 1$ , and the proof is now complete.  $\square$

Note that the requirement  $\|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\| < 1$  gives a restriction on the admissible choices of  $\varepsilon(\mathbf{x})$ , as we see in the following example.

**Example 3.11. (Example 3.9 continued)**

Let  $\varepsilon(\mathbf{x}) = \{\delta_1, \delta_2 y^2 + \delta_3\}$ , with  $\delta_i \in \mathbb{R}$ , be an admissible perturbation of the zero-dimensional complete intersection  $\mathbf{f}(\mathbf{x})$  of Example 3.9. We consider the real solution  $p_4 = (2, 3)$  of  $\mathbf{f} = 0$  and compute  $\|\text{Jac}_{\mathbf{f}}(p_4)^{-1} \text{Jac}_{\varepsilon}(p_4)\|_2^2 = \frac{117}{25} \delta_2^2$ . From Lemma 3.10 the condition  $|\delta_2| < \frac{5}{39} \sqrt{13}$  is sufficient to have  $\text{Jac}_{\mathbf{f}+\varepsilon}(p_4)$  invertible.

From now on we assume that the hypothesis of Lemma 3.10 is satisfied. In order to deduce an upper bound for  $\|\Delta p\|$  we consider an approximation of it.

**Definition 3.12.** If  $\|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\|$  is different from 1, we denote the number  $1/(1 - \|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\|)$  by  $\Lambda(\mathbf{f}, \varepsilon, p)$ . Moreover, if equation (1) is truncated at the first order, we get the approximate solution  $-\text{Jac}_{\mathbf{f}+\varepsilon}(p)^{-1} \varepsilon(p)$  which we call  $\Delta p^1$ .

**Proposition 3.13.** *Assume that  $\|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\| < 1$  and let  $\|\cdot\|$  be an induced matrix norm. Then we have*

$$(2) \quad \|\Delta p^1\| \leq \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_{\mathbf{f}}(p)^{-1}\| \|\varepsilon(p)\|$$

*Proof.* Lemma 3.10 guarantees that the matrix  $\text{Jac}_{\mathbf{f}+\varepsilon}(p)$  is invertible, so

$$\begin{aligned} \Delta p^1 &= -\text{Jac}_{\mathbf{f}+\varepsilon}(p)^{-1} \varepsilon(p) = -(\text{Jac}_{\mathbf{f}}(p) + \text{Jac}_{\varepsilon}(p))^{-1} \varepsilon(p) \\ &= -(I + \text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p))^{-1} \text{Jac}_{\mathbf{f}}(p)^{-1} \varepsilon(p) \end{aligned}$$

We apply the inequality of Proposition 3.2 to  $\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)$ , and get

$$\begin{aligned} \|\Delta p^1\| &\leq \|(I + \text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p))^{-1}\| \|\text{Jac}_{\mathbf{f}}(p)^{-1}\| \|\varepsilon(p)\| \\ &\leq \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_{\mathbf{f}}(p)^{-1}\| \|\varepsilon(p)\| \end{aligned}$$

which concludes the proof.  $\square$

We introduce the local condition number of the polynomial system  $\mathbf{f}(\mathbf{x}) = 0$ .

**Definition 3.14.** Let  $\mathbf{f}(\mathbf{x})$  be a zero-dimensional smooth complete intersection in  $\mathbb{R}[\mathbf{x}]$ , let  $p$  be a nonsingular real solution of  $\mathbf{f}(\mathbf{x}) = 0$ , and let  $\|\cdot\|$  be a norm.

- (a) The number  $\kappa(\mathbf{f}, p) = \|\text{Jac}_{\mathbf{f}}(p)^{-1}\| \|\text{Jac}_{\mathbf{f}}(p)\|$  is called the **local condition number** of  $\mathbf{f}(\mathbf{x})$  at  $p$ .
- (b) If the norm is an  $r$ -norm, the local condition number is denoted by  $\kappa_r(\mathbf{f}, p)$ .

The following theorem illustrates the importance of the local condition number. It depends on  $f$  and  $p$ , not on  $\varepsilon$  and is a key ingredient to provide an upper bound for the relative error  $\frac{\|\Delta p^1\|}{\|p\|}$ .

**Theorem 3.15. (Local Condition Number)**

Let  $\|\cdot\|$  be an induced matrix norm; under the above assumptions and the condition  $\|\text{Jac}_f(p)^{-1} \text{Jac}_\varepsilon(p)\| < 1$  we have

$$(3) \quad \frac{\|\Delta p^1\|}{\|p\|} \leq \Lambda(\mathbf{f}, \varepsilon, p) \kappa(\mathbf{f}, p) \left( \frac{\|\text{Jac}_\varepsilon(p)\|}{\|\text{Jac}_f(p)\|} + \frac{\|\varepsilon(0) - \varepsilon^{\geq 2}(0, p)\|}{\|\mathbf{f}(0) - \mathbf{f}^{\geq 2}(0, p)\|} \right)$$

*Proof.* By Definition 3.5 the evaluation of  $\varepsilon$  at 0 can be expressed in this way  $\varepsilon(0) = \varepsilon(p) - \text{Jac}_\varepsilon(p)p + \varepsilon^{\geq 2}(0, p)$ , and so  $\varepsilon(p) = \varepsilon(0) + \text{Jac}_\varepsilon(p)p - \varepsilon^{\geq 2}(0, p)$ . Dividing (2) of Proposition 3.13 by  $\|p\|$  we obtain

$$\begin{aligned} \frac{\|\Delta p^1\|}{\|p\|} &\leq \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_f(p)^{-1}\| \frac{\|\varepsilon(p)\|}{\|p\|} \\ &\leq \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_f(p)^{-1}\| \frac{\|\text{Jac}_\varepsilon(p)\| \|p\| + \|\varepsilon(0) - \varepsilon^{\geq 2}(0, p)\|}{\|p\|} \\ &= \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_f(p)^{-1}\| \left( \|\text{Jac}_\varepsilon(p)\| + \frac{\|\varepsilon(0) - \varepsilon^{\geq 2}(0, p)\|}{\|p\|} \right) \end{aligned}$$

Using again Definition 3.5 we express  $\mathbf{f}(0) = \mathbf{f}(p) - \text{Jac}_f(p)p + \mathbf{f}^{\geq 2}(0, p)$ ; since  $\mathbf{f}(p) = 0$  we have  $\|\mathbf{f}(0) - \mathbf{f}^{\geq 2}(0, p)\| = \|\text{Jac}_f(p)p\| \leq \|\text{Jac}_f(p)\| \|p\|$  from which

$$\frac{1}{\|p\|} \leq \frac{\|\text{Jac}_f(p)\|}{\|\mathbf{f}(0) - \mathbf{f}^{\geq 2}(0, p)\|}$$

We combine the inequalities to obtain

$$\begin{aligned} \frac{\|\Delta p^1\|}{\|p\|} &\leq \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_f(p)^{-1}\| \left( \|\text{Jac}_\varepsilon(p)\| + \|\text{Jac}_f(p)\| \frac{\|\varepsilon(0) - \varepsilon^{\geq 2}(0, p)\|}{\|\mathbf{f}(0) - \mathbf{f}^{\geq 2}(0, p)\|} \right) \\ &\leq \Lambda(\mathbf{f}, \varepsilon, p) \|\text{Jac}_f(p)^{-1}\| \|\text{Jac}_f(p)\| \left( \frac{\|\text{Jac}_\varepsilon(p)\|}{\|\text{Jac}_f(p)\|} + \frac{\|\varepsilon(0) - \varepsilon^{\geq 2}(0, p)\|}{\|\mathbf{f}(0) - \mathbf{f}^{\geq 2}(0, p)\|} \right) \end{aligned}$$

and the proof is concluded.  $\square$

The following remark contains observations about the local condition number.

**Remark 3.16.** We call attention to the following observations.

- (a) The notion of local condition number given in Definition 3.14 is a generalization of the classical notion of condition number of linear systems (see [4]). In fact, if  $\mathbf{f}(\mathbf{x})$  is linear, that is  $\mathbf{f}(\mathbf{x}) = A\mathbf{x} - b$  with  $A \in \text{Mat}_n(\mathbb{R})$  invertible, and  $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}) = \{p\} = \{A^{-1}b\}$ , then  $\kappa(\mathbf{f}, p)$  is the classical condition number of the matrix  $A$ . In fact  $\text{Jac}_f(\mathbf{x}) = A$ , and so  $\kappa(\mathbf{f}, p) = \|\text{Jac}_f(p)^{-1}\| \|\text{Jac}_f(p)\| = \|A^{-1}\| \|A\|$ . Further, if we consider the perturbation  $\varepsilon(\mathbf{x}) = \Delta A\mathbf{x} - \Delta b$ , relation (3) becomes

$$\frac{\|\Delta p\|}{\|p\|} \leq \frac{1}{1 - \|A^{-1}\| \|\Delta A\|} \|A^{-1}\| \|A\| \left( \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right) \quad (4)$$

which is the relation that quantifies the sensitivity of the  $Ax = b$  problem (see [4], Theorem 4.1).

- (b) Using any induced matrix norm, the condition number  $\kappa(\mathbf{f}, p)$  turns out to be greater than or equal to 1. In particular, using the 2-norm we have  $\kappa_2(\mathbf{f}, p) = \frac{\sigma_{\max}(\text{Jac}_{\mathbf{f}}(p))}{\sigma_{\min}(\text{Jac}_{\mathbf{f}}(p))}$ ; in this case the local condition number attains its minimum, that is  $\kappa_2(\mathbf{f}, p) = 1$ , when  $\text{Jac}_{\mathbf{f}}(p)$  is orthonormal.
- (c) The condition number  $\kappa(\mathbf{f}, p)$  is invariant under a scalar multiplication of the polynomial system  $\mathbf{f}(\mathbf{x})$  by a unique nonzero real number  $\gamma$ . On the contrary,  $\kappa(\mathbf{f}, p)$  is not invariant under a generic scalar multiplication of each polynomial  $f_j(\mathbf{x})$  of  $\mathbf{f}(\mathbf{x})$ . The reason is that if we multiply each  $f_j(\mathbf{x})$  by a nonzero real number  $\gamma_j$  we obtain the new polynomial set  $\mathbf{g}(\mathbf{x}) = \{\gamma_1 f_1(\mathbf{x}), \dots, \gamma_n f_n(\mathbf{x})\}$  whose condition number at  $p$  is

$$\kappa(\mathbf{g}, p) = \|\text{Jac}_{\mathbf{f}}(p)^{-1} \Gamma^{-1}\| \|\Gamma \text{Jac}_{\mathbf{f}}(p)\| \neq \kappa(\mathbf{f}, p)$$

where  $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_n) \in \text{Mat}_n(\mathbb{R})$  is the diagonal matrix with entries  $\gamma_1, \dots, \gamma_n$ .

- (d) It is interesting to observe that if  $p$  is the origin then Formula (3) of the theorem is not applicable. However, one can translate  $p$  away from the origin, and the nice thing is that the local condition number does not change.

#### 4. OPTIMIZATION OF THE LOCAL CONDITION NUMBER

In this section we introduce a strategy to improve the numerical stability of zero-dimensional smooth complete intersections. Let  $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$  be a zero-dimensional smooth complete intersection in  $\mathbb{R}[\mathbf{x}]$ , and let  $I$  be the ideal of  $\mathbb{R}[\mathbf{x}]$  generated by  $\mathbf{f}(\mathbf{x})$ ; our aim is to find an alternative representation of  $I$  with minimal local condition number.

Motivated by Remark 3.16, item (b) and (c), we consider the strategy of resizing each polynomial of  $\mathbf{f}(\mathbf{x})$ , and study its effects on the condition number. The following proposition shows that rescaling each  $f_j(\mathbf{x})$  so that  $\frac{\partial f_j}{\partial \mathbf{x}}(p)$  has unitary norm is a nearly optimal, in some cases optimal, strategy. The result is obtained by adapting the method of Van der Sluis (see [13], Section 7.3) to the polynomial case.

**Proposition 4.1.** *Let  $p$  be a nonsingular real solution of  $\mathbf{f}(\mathbf{x}) = 0$ , let  $r_1 \geq 1, r_2 \geq 1$  be real numbers such that  $\frac{1}{r_1} + \frac{1}{r_2} = 1$ , including the pairs  $(1, \infty)$  and  $(\infty, 1)$ , let  $\gamma = (\gamma_1, \dots, \gamma_n)$  be an  $n$ -tuple of nonzero real numbers, and let  $\mathbf{g}_{\gamma}(\mathbf{x})$ ,  $\mathbf{u}(\mathbf{x})$  be the polynomial systems defined by  $\mathbf{g}_{\gamma}(\mathbf{x}) = \{\gamma_1 f_1(\mathbf{x}), \dots, \gamma_n f_n(\mathbf{x})\}$  and  $\mathbf{u}(\mathbf{x}) = \{\|\frac{\partial f_1}{\partial \mathbf{x}}(p)\|_{r_2}^{-1} f_1(\mathbf{x}), \dots, \|\frac{\partial f_n}{\partial \mathbf{x}}(p)\|_{r_2}^{-1} f_n(\mathbf{x})\}$ .*

- (a) *We have the inequality  $\kappa_{r_1}(\mathbf{u}, p) \leq n^{1/r_1} \kappa_{r_1}(\mathbf{g}_{\gamma}, p)$ .*
- (b) *In particular, if  $(r_1, r_2) = (\infty, 1)$  we have the equality*

$$\kappa_{\infty}(\mathbf{u}, p) = \min_{\gamma} \kappa_{\infty}(\mathbf{g}_{\gamma}, p)$$

where  $\mathbf{u}(\mathbf{x}) = \{\|\frac{\partial f_1}{\partial \mathbf{x}}(p)\|_1^{-1} f_1(\mathbf{x}), \dots, \|\frac{\partial f_n}{\partial \mathbf{x}}(p)\|_1^{-1} f_n(\mathbf{x})\}$ .

*Proof.* Let  $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_n)$  and  $D = \text{diag}(\|\frac{\partial f_1}{\partial \mathbf{x}}(p)\|_{r_2}^{-1}, \dots, \|\frac{\partial f_n}{\partial \mathbf{x}}(p)\|_{r_2}^{-1})$ ; then  $\text{Jac}_{\mathbf{g}_{\gamma}}(\mathbf{x}) = \Gamma \text{Jac}_{\mathbf{f}}(\mathbf{x})$  and  $\text{Jac}_{\mathbf{u}}(\mathbf{x}) = D \text{Jac}_{\mathbf{f}}(\mathbf{x})$ . The condition numbers of  $\mathbf{g}_{\gamma}(\mathbf{x})$  and  $\mathbf{u}(\mathbf{x})$  at  $p$  are given by

$$\begin{aligned} \kappa_{r_1}(\mathbf{g}_{\gamma}, p) &= \|(\Gamma \text{Jac}_{\mathbf{f}}(p))^{-1}\|_{r_1} \|\Gamma \text{Jac}_{\mathbf{f}}(p)\|_{r_1} \\ \kappa_{r_1}(\mathbf{u}, p) &= \|(D \text{Jac}_{\mathbf{f}}(p))^{-1}\|_{r_1} \|D \text{Jac}_{\mathbf{f}}(p)\|_{r_1} \end{aligned}$$

From Proposition 3.3 we have

$$\begin{aligned}
\|D \text{Jac}_{\mathbf{f}}(p)\|_{r_1} &\leq n^{1/r_1} \max_i \|(D \text{Jac}_{\mathbf{f}}(p))_i\|_{r_2} = n^{1/r_1} \\
\|(D \text{Jac}_{\mathbf{f}}(p))^{-1}\|_{r_1} &= \|\text{Jac}_{\mathbf{f}}^{-1}(p) D^{-1}\|_{r_1} = \|\text{Jac}_{\mathbf{f}}^{-1}(p) \Gamma^{-1} \Gamma D^{-1}\|_{r_1} \\
&\leq \|\text{Jac}_{\mathbf{f}}^{-1}(p) \Gamma^{-1}\|_{r_1} \max_i \left( |\gamma_i| \left\| \frac{\partial f_i}{\partial \mathbf{x}}(p) \right\|_{r_2} \right) \\
&\leq \|\text{Jac}_{\mathbf{f}}^{-1}(p) \Gamma^{-1}\|_{r_1} \|\Gamma \text{Jac}_{\mathbf{f}}(p)\|_{r_1} = \kappa_{r_1}(\mathbf{g}_{\gamma}, p)
\end{aligned}$$

therefore  $\kappa_{r_1}(\mathbf{u}, p) \leq n^{1/r_1} \kappa_{r_1}(\mathbf{g}_{\gamma}, p)$  and (a) is proved. To prove (b) it suffices to use (a) and observe that  $n^{1/\infty} = 1$   $\square$

**Remark 4.2.** The above proposition implies that the strategy of rescaling each polynomial  $f_j(\mathbf{x})$  to make it unitary at  $p$  (see Definition 3.5) is beneficial for lowering the local condition number of  $\mathbf{f}(\mathbf{x})$  at  $p$ . This number is minimum when  $r = \infty$ , it is within factor  $\sqrt{n}$  of the minimum when  $r = 2$ . However, for  $r = 2$  we can do better, at least when all the polynomials  $f_1(\mathbf{x}), \dots, f_n(\mathbf{x})$  have equal degree. The idea is to use Remark 3.16, item (b) which says that when using the matrix 2-norm, the local condition number attains its minimum when the Jacobian matrix is orthonormal.

**Proposition 4.3.** *Let  $\mathbf{f} = (f_1, \dots, f_n)$  be a smooth zero-dimensional complete intersection in  $\mathbb{R}[\mathbf{x}]$  such that  $\deg(f_1) = \dots = \deg(f_n)$  and let  $p \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ . Moreover, let  $C = (c_{ij}) \in \text{Mat}_n(\mathbb{R})$  be an invertible matrix, and let  $\mathbf{g}$  be defined by  $\mathbf{g}^{\text{tr}} = C \cdot \mathbf{f}^{\text{tr}}$ . Then the following conditions are equivalent*

- (a)  $\kappa_2(\mathbf{g}, p) = 1$ , the minimum possible.
- (b)  $C^t C = (\text{Jac}_{\mathbf{f}}(p) \text{Jac}_{\mathbf{f}}(p)^t)^{-1}$ .

*Proof.* We know that  $\kappa_2(\mathbf{g}, p) = 1$  if and only if the matrix  $\text{Jac}_{\mathbf{g}}(p)$  is orthonormal. This condition can be expressed by the equality  $\text{Jac}_{\mathbf{g}}(p) \text{Jac}_{\mathbf{g}}(p)^t = I_n$ , that is  $C \text{Jac}_{\mathbf{f}}(p) \text{Jac}_{\mathbf{f}}(p)^t C^t = I_n$  and the conclusion follows.  $\square$

We observe that condition (b) of the proposition requires that the entries of  $C$  satisfy an underdetermined system of  $(n^2 + n)/2$  independent quadratic equations in  $n^2$  unknowns.

## 5. EXPERIMENTS

In numerical linear algebra it is well-known (see for instance [4], Ch. 4, Section 1) that the upper bound given by the classical formula (4) of Remark 3.16 (a) is not necessarily sharp. Since our upper bound (3) generalizes the classical one, as shown in Remark 3.16, we provide some experimental evidence that lowering the condition number not only sharpens the upper bound, but indeed stabilizes the solution point.

**Example 5.1.** We consider the ideal  $I = (f_1, f_2)$  in  $\mathbb{R}[x, y]$  where

$$\begin{aligned}
f_1 &= \frac{1}{4}x^2y + xy^2 + \frac{1}{4}y^3 + \frac{1}{5}x^2 - \frac{5}{8}xy + \frac{13}{40}y^2 + \frac{9}{40}x - \frac{3}{5}y + \frac{1}{40} \\
f_2 &= x^3 + \frac{14}{13}xy^2 + \frac{57}{52}x^2 - \frac{25}{52}xy + \frac{8}{13}y^2 - \frac{11}{52}x - \frac{4}{13}y - \frac{4}{13}
\end{aligned}$$

It is a zero-dimensional smooth complete intersection with 7 real roots and we consider the point  $p = (0, 1) \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ . The polynomial system  $\mathbf{f} = \{f_1, f_2\}$  is unitary at  $p$  and its condition number is  $\kappa_2(\mathbf{f}, p) = 8$ . Using Proposition 4.3 we construct a new polynomial system  $\mathbf{g}$  with minimal local condition number at  $p$ .



The new pair of generators  $\mathbf{g}$  is defined (see Proposition 4.3) by the following the formula  $\mathbf{g}^{\text{tr}} = C \cdot \mathbf{f}^{\text{tr}}$ , where  $C = (c_{ij}) \in \text{Mat}_2(\mathbb{R})$  is an invertible matrix whose entries satisfy the following system

$$\begin{cases} c_{11}^2 + c_{21}^2 &= \frac{25}{16} \\ c_{11}c_{12} + c_{21}c_{22} &= -\frac{15}{16} \\ c_{12}^2 + c_{22}^2 &= \frac{25}{16} \end{cases}$$

A solution is given by  $c_{11} = 1$ ,  $c_{12} = 0$ ,  $c_{21} = \frac{63}{16}$ ,  $c_{22} = -\frac{65}{16}$ , and we observe that the associated unitary polynomial system  $\mathbf{g} = \{f_1, \frac{63}{16}f_1 - \frac{65}{16}f_2\}$  provides an alternative representation of  $I$  with minimal local condition number  $\kappa_2(\mathbf{g}, p) = 1$  at the point  $p$ .

Now we embed the system  $\mathbf{f}(x, y)$  into the family  $F(a, x, y) = \{F_1, F_2\}$  where

$$\begin{aligned} F_1(a, x, y) &= \frac{1}{4}x^2y + xy^2 + \frac{1}{4}y^3 + \frac{1}{5}x^2 - \frac{5}{8}xy + \left(\frac{13}{40} - a\right)y^2 \\ &\quad + \left(\frac{9}{40} + a\right)x + \left(-\frac{3}{5} + a\right)y + \frac{1}{40} - 2a \\ F_2(a, x, y) &= x^3 + \frac{14}{13}xy^2 + \frac{57}{52}x^2 - \frac{25}{52}xy + \left(\frac{8}{13} + a\right)y^2 \\ &\quad + \left(-\frac{11}{52} + a\right)x - \left(\frac{4}{13} + a\right)y - \frac{4}{13} + a^2 \end{aligned}$$

We denote by  $I_F(a, x, y)$  the ideal generated by  $F(a, x, y)$  in  $\mathbb{R}[a, x, y]$ , compute the reduced **Lex**-Gröbner basis of  $I_F(a, x, y)\mathbb{R}(a)[x, y]$ , and get

$$\left\{x + \frac{l_1(a, y)}{d_F(a)}, y^9 + l_2(a, y)\right\}$$

where  $l_1(a, y), l_2(a, y) \in \mathbb{R}[a, y]$  have degree 8 in  $y$  and  $d_F(a) \in \mathbb{R}[a]$  has degree 12. This basis has the shape prescribed by the Shape Lemma and a flat locus is given by  $\{\alpha \in \mathbb{R} \mid d_F(\alpha) \neq 0\}$ . We let  $D_F(a, x, y) = \det(\text{Jac}_F(a, x, y))$ ,  $J_F(a, x, y) = I_F(a, x, y) + (D_F(a, x, y))$ , compute  $J_F(a, x, y) \cap \mathbb{R}[a]$ , and we get the principal ideal generated by a univariate polynomial  $h_F(a)$  of degree 28. An  $I$ -optimal subscheme is  $\mathcal{U}_F = \{\alpha \in \mathbb{R} \mid d_F(\alpha)h_F(\alpha) \neq 0\}$ . An open semi-algebraic subset  $\mathcal{V}_F$  of  $\mathcal{U}_F$  which contains the point  $\alpha_I = 0$  and such that the fiber over each  $\alpha \in \mathcal{V}_F$  consists of 7 real points, is given by the open interval  $(\alpha_1, \alpha_2)$ , where  $\alpha_1 < 0$  and  $\alpha_2 > 0$  are the real roots of  $d_F(a)h_F(a) = 0$  closest to the origin. Their approximate values are  $\alpha_1 = -0.00006$  and  $\alpha_2 = 0.01136$ .

To produce similar perturbations, we embed the system  $\mathbf{g}(x, y)$  into the family  $G(a, x, y) = \{G_1, G_2\}$  where

$$\begin{aligned} G_1(a, x, y) &= \frac{1}{4}x^2y + xy^2 + \frac{1}{4}y^3 + \frac{1}{5}x^2 - \frac{5}{8}xy + \left(\frac{13}{40} - a\right)y^2 \\ &\quad + \left(\frac{9}{40} + a\right)x + \left(-\frac{3}{5} + a\right)y + \frac{1}{40} - 2a \\ G_2(a, x, y) &= -\frac{65}{16}x^3 + \frac{63}{64}x^2y - \frac{7}{16}xy^2 + \frac{63}{64}y^3 - \frac{1173}{320}x^2 - \frac{65}{128}xy \\ &\quad + \left(-\frac{781}{640} + a\right)y^2 + \left(\frac{1117}{640} + a\right)x + \left(-\frac{89}{80} - a\right)y + \frac{863}{640} + a^2 \end{aligned}$$

We denote by  $I_G(a, x, y)$  the ideal generated by  $G(a, x, y)$  in  $\mathbb{R}[a, x, y]$ , compute the reduced **Lex**-Gröbner basis of  $I_G(a, x, y)\mathbb{R}(a)[x, y]$ , and get

$$\left\{x + \frac{l_3(a, y)}{d_G(a)}, y^9 + l_4(a, y)\right\}$$

where  $l_3(a, y), l_4(a, y) \in \mathbb{R}[a, y]$  have degree 8 in  $y$  and  $d_G(a) \in \mathbb{R}[a]$  has degree 12, therefore the basis has the shape prescribed by the Shape Lemma. A flat locus is given by  $\{\alpha \in \mathbb{R} \mid d_G(\alpha) \neq 0\}$ . We let  $D_G(a, x, y) = \det(\text{Jac}_G(a, x, y))$ ,  $J_G(a, x, y) = I_G(a, x, y) + (D_G(a, x, y))$  and compute  $J_G(a, x, y) \cap \mathbb{R}[a]$ . We get the principal ideal generated by a univariate polynomial  $h_G(a)$  of degree 28. An

$I$ -optimal subscheme is  $\mathcal{U}_G = \{\alpha \in \mathbb{R} \mid d_G(\alpha)h_G(\alpha) \neq 0\}$ . An open semi-algebraic subset  $\mathcal{V}_G$  of  $\mathcal{U}_G$  containing the point  $\alpha_I = 0$  and such that the fiber over each  $\alpha \in \mathcal{V}_G$  consists of 7 real points is given by the open interval  $(\alpha_3, \alpha_4)$ , where  $\alpha_3 < 0$  and  $\alpha_4 > 0$  are the real roots of  $d_G(a)h_G(a) = 0$  closest to the origin. Their approximate values are  $\alpha_3 = -0.00009$  and  $\alpha_4 = 0.00914$ .

Let  $\alpha \in (\alpha_1, \alpha_4)$ . According to Definition 3.7 the polynomial set  $\varepsilon(x, y) = \{-\alpha y^2 + \alpha x + \alpha y - 2\alpha, \alpha y^2 + \alpha x - \alpha y + \alpha^2\}$  is an admissible perturbation of  $\mathbf{f}(x, y)$  and  $\mathbf{g}(x, y)$ . Further, since  $\|\text{Jac}_{\mathbf{f}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\|_2 = \sqrt{65}|\alpha| < 1$  and  $\|\text{Jac}_{\mathbf{g}}(p)^{-1} \text{Jac}_{\varepsilon}(p)\|_2 = \sqrt{2}|\alpha| < 1$  Theorem 3.15 can be applied.

We let  $q \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \varepsilon)$  and  $r \in \mathcal{Z}_{\mathbb{R}}(\mathbf{g} + \varepsilon)$  be the two perturbations of the point  $p$ . In order to compare the numerical behaviour of  $\mathbf{f}$  and  $\mathbf{g}$  at the real root  $p$  we compare the relative errors  $\frac{\|q-p\|_2}{\|p\|_2}$  and  $\frac{\|r-p\|_2}{\|p\|_2}$  for different values of  $\alpha$ . The first column of the following table contains the values of the local condition numbers of  $\mathbf{f}$  and  $\mathbf{g}$  at  $p$ . The second column contains the mean values of the upper bounds  $\text{UB}(\mathbf{f}, p)$  and  $\text{UB}(\mathbf{g}, p)$  given by Theorem 3.15, computed for 100 random values of  $\alpha \in (\alpha_1, \alpha_4)$ . The third column contains the mean values of  $\frac{\|q-p\|_2}{\|p\|_2}$  and  $\frac{\|r-p\|_2}{\|p\|_2}$  for the same values of  $\alpha$ .

$\kappa_2(\mathbf{f}, p)$	$\text{UB}(\mathbf{f}, p)$	$\frac{\ q-p\ _2}{\ p\ _2}$
8	0.1729	0.000097
$\kappa_2(\mathbf{g}, p)$	$\text{UB}(\mathbf{g}, p)$	$\frac{\ r-p\ _2}{\ p\ _2}$
1	0.0275	0.000023

The fact that the mean values of  $\frac{\|q-p\|_2}{\|p\|_2}$  are smaller than the mean values of  $\frac{\|r-p\|_2}{\|p\|_2}$  suggests that  $p$  is more stable when it is considered as a root of  $\mathbf{g}$  instead of as a root of  $\mathbf{f}$ .

**Example 5.2.** We consider the ideal  $I = (f_1, f_2, f_3)$  in  $\mathbb{R}[x, y, z]$  where

$$\begin{aligned} f_1 &= \frac{6}{17}x^2 + xy - \frac{24}{85}x - \frac{8}{85}y - \frac{6}{85} \\ f_2 &= \frac{39}{89}x^2 + \frac{70}{89}xy + yz - \frac{39}{89}x + \frac{10}{89}y \\ f_3 &= y^2 + 2xz + z^2 - z \end{aligned}$$

It is a zero-dimensional smooth complete intersection with 6 real roots and we consider the point  $p = (1, 0, 0) \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ . The polynomial system  $\mathbf{f} = \{f_1, f_2, f_3\}$  is unitary at  $p$  and its condition number is  $\kappa_2(\mathbf{f}, p) = 123$ . Using Proposition 4.3 we construct a new polynomial system  $\mathbf{g}$  with minimal local condition number at  $p$ . The new set  $\mathbf{g}$  is defined by  $\mathbf{g}^{\text{tr}} = C \cdot \mathbf{f}^{\text{tr}}$ , where  $C = (c_{ij}) \in \text{Mat}_3(\mathbb{R})$  is an invertible matrix whose entries satisfy the following system

$$\begin{cases} c_{11}^2 + c_{21}^2 + c_{31}^2 &= \frac{57229225}{15129} \\ c_{11}c_{12} + c_{21}c_{22} + c_{31}c_{32} &= -\frac{57221660}{15129} \\ c_{11}c_{13} + c_{21}c_{23} + c_{31}c_{33} &= 0 \\ c_{12}^2 + c_{22}^2 + c_{32}^2 &= \frac{57229225}{15129} \\ c_{12}c_{13} + c_{22}c_{23} + c_{32}c_{33} &= 0 \\ c_{13}^2 + c_{23}^2 + c_{33}^2 &= 1 \end{cases}$$

A solution is given by  $c_{11} = c_{33} = 1$ ,  $c_{12} = c_{13} = c_{23} = c_{32} = 0$ ,  $c_{21} = \frac{7564}{123}$ ,  $c_{22} = -\frac{7565}{123}$ . Therefore the associated unitary polynomial system is the following

$\mathbf{g} = \{f_1, \frac{7564}{123}f_1 - \frac{7565}{123}f_2, f_3\}$ . It provides an alternative representation of  $I$  with minimal local condition number  $\kappa_2(\mathbf{g}, p) = 1$  at the point  $p$ .

We embed the system  $\mathbf{f}(x, y, z)$  into the family  $F(a, x, y, z) = \{F_1, F_2, F_3\}$  where

$$\begin{aligned} F_1(a, x, y, z) &= \frac{6}{17}x^2 + (1 - a^2)xy + (-\frac{24}{85} + a)x + (-\frac{8}{85} - a)y + (-\frac{6}{85} + a^2) \\ F_2(a, x, y, z) &= \frac{39}{89}x^2 + (\frac{70}{89} + a)xy + yz + (\frac{39}{89} + a)x + (\frac{10}{89} + a)y \\ F_3(a, x, y, z) &= y^2 + 2xz + (1 - 2a)z^2 + (-1 + a)z \end{aligned}$$

We denote by  $I_F(a, x, y, z)$  the ideal generated by  $F(a, x, y, z)$  in  $\mathbb{R}[a, x, y, z]$ , compute the reduced **Lex**-Gröbner basis of  $I_F(a, x, y, z)\mathbb{R}(a)[x, y, z]$ , and get

$$\{x + \frac{l_1(a, z)}{d_F(a)}, y + \frac{l_2(a, z)}{d_F(a)}, z^9 + \frac{l_3(a, z)}{e_F(a)}\}$$

where  $l_1(a, z), l_2(a, z), l_3(a, z) \in \mathbb{R}[a, z]$  have degrees  $\deg_z(l_1) = \deg_z(l_2) = 7$  and  $\deg_z(l_3) = 8$  while  $d_F(a) \in \mathbb{R}[a]$  has degree 54, and  $e_F(a) \in \mathbb{R}[a]$  has degree 11. The basis has the shape prescribed by the Shape Lemma. A flat locus is given by  $\{\alpha \in \mathbb{R} \mid d_F(\alpha)e_F(\alpha) \neq 0\}$ . We let  $D_F(a, x, y, z) = \det(\text{Jac}_F(a, x, y, z))$ ,  $J_F(a, x, y, z) = I_F(a, x, y, z) + (D_F(a, x, y, z))$  and compute  $J_F(a, x, y, z) \cap \mathbb{R}[a]$ . We get the principal ideal generated by a univariate polynomial  $h_F(a)$  of degree 59. An  $I$ -optimal subscheme is  $\mathcal{U}_F = \{\alpha \in \mathbb{R} \mid d_F(\alpha)e_F(\alpha)h_F(\alpha) \neq 0\}$ . An open semi-algebraic subset  $\mathcal{V}_F$  of  $\mathcal{U}_F$  containing the point  $\alpha_I = 0$  and such that the fiber over each  $\alpha \in \mathcal{V}_F$  consists of 6 real points is given by the open interval  $(\alpha_1, \alpha_2)$ , where  $\alpha_1 < 0$  and  $\alpha_2 > 0$  are the real roots of  $d_F(a)e_F(a)h_F(a) = 0$  closest to the origin. Their approximate values are  $\alpha_1 = -0.17082$  and  $\alpha_2 = 0.20711$ .

To produce similar perturbations, we embed the system  $\mathbf{g}(x, y, z)$  into the family  $G(a, x, y, z) = \{G_1, G_2, G_3\}$  where

$$\begin{aligned} G_1(a, x, y) &= \frac{6}{17}x^2 + (1 - a^2)xy + (-\frac{24}{85} + a)x + (-\frac{8}{85} - a)y + (-\frac{6}{85} + a^2) \\ G_2(a, x, y) &= -\frac{3657}{697}x^2 + (\frac{538}{41} + a)xy - \frac{7565}{123}yz + (\frac{33413}{3485} + a)x \\ &\quad + (-\frac{44254}{3485} + a)y - \frac{15128}{3485} \\ G_3(a, x, y) &= y^2 + 2xz + (1 - 2a)z^2 + (-1 + a)z \end{aligned}$$

We denote by  $I_G(a, x, y, z)$  the ideal generated by  $G(a, x, y, z)$  in  $\mathbb{R}[a, x, y, z]$ , compute the reduced **Lex**-Gröbner basis of  $I_G(a, x, y, z)\mathbb{R}(a)[x, y, z]$ , and get

$$\{x + \frac{l_4(a, z)}{d_G(a)}, y + \frac{l_5(a, z)}{d_G(a)}, z^9 + \frac{l_6(a, z)}{e_G(a)}\}$$

where  $l_4(a, z), l_5(a, z), l_6(a, z) \in \mathbb{R}[a, z]$  have degrees  $\deg_z(l_4) = \deg_z(l_5) = 7$  and  $\deg_z(l_6) = 8$  while  $d_G(a) \in \mathbb{R}[a]$  has degree 54, and  $e_G(a) \in \mathbb{R}[a]$  has degree 11. The basis has the shape prescribed by the Shape Lemma. A flat locus is given by  $\{\alpha \in \mathbb{R} \mid d_{G1}(\alpha)d_{G2}(\alpha) \neq 0\}$ . We let  $D_G(a, x, y, z) = \det(\text{Jac}_G(a, x, y, z))$ ,  $J_G(a, x, y, z) = I_G(a, x, y, z) + (D_G(a, x, y, z))$  and compute  $J_G(a, x, y, z) \cap \mathbb{R}[a]$ . We get the principal ideal generated by a univariate polynomial  $h_G(a)$  of degree 59. An  $I$ -optimal subscheme is  $\mathcal{U}_G = \{\alpha \in \mathbb{R} \mid d_G(\alpha)e_G(\alpha)h_G(\alpha) \neq 0\}$ . An open semi-algebraic subset  $\mathcal{V}_G$  of  $\mathcal{U}_G$  containing the point  $\alpha_I = 0$  and such that the fiber over each  $\alpha \in \mathcal{V}_G$  consists of 6 real points is given by the open interval  $(\alpha_3, \alpha_4)$ , where  $\alpha_3 < 0$  and  $\alpha_4 > 0$  are the real roots of  $d_G(a)e_G(a)h_G(a) = 0$  closest to the origin. Their approximate values are  $\alpha_3 = -0.02942$  and  $\alpha_4 = 0.03312$ .

Let  $\alpha \in (\alpha_3, \alpha_4)$ . According to Definition 3.7 the polynomial set  $\varepsilon(x, y) = \{-\alpha^2xy + \alpha x - \alpha y + \alpha^2, \alpha xy + \alpha x + \alpha y, -2\alpha z^2 + \alpha z\}$  is an admissible perturbation of  $\mathbf{f}(x, y, z)$  and  $\mathbf{g}(x, y, z)$ .

We let  $q \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \varepsilon)$  and  $r \in \mathcal{Z}_{\mathbb{R}}(\mathbf{g} + \varepsilon)$  be the two perturbations of the point  $p$ . In order to compare the numerical behaviour of  $\mathbf{f}$  and  $\mathbf{g}$  at the real root  $p$  we compare the relative errors  $\frac{\|q-p\|_2}{\|p\|_2}$  and  $\frac{\|r-p\|_2}{\|p\|_2}$  for different values of  $\alpha$ . The first column of the following table contains the values of the local condition numbers of  $\mathbf{f}$  and  $\mathbf{g}$  at  $p$ . The second column contains the mean values of  $\frac{\|q-p\|_2}{\|p\|_2}$  and  $\frac{\|r-p\|_2}{\|p\|_2}$  for 100 random values of  $\alpha \in (\alpha_1, \alpha_4)$ .

$\kappa_2(\mathbf{f}, p)$	$\frac{\ q-p\ _2}{\ p\ _2}$
123	0.0436
$\kappa_2(\mathbf{g}, p)$	$\frac{\ r-p\ _2}{\ p\ _2}$
1	0.0221

As in the example before, the fact that the mean values of  $\frac{\|q-p\|_2}{\|p\|_2}$  are smaller than the mean values of  $\frac{\|r-p\|_2}{\|p\|_2}$  suggests that  $p$  is more stable when it is considered as a root of  $\mathbf{g}$  instead of as a root of  $\mathbf{f}$ .

## REFERENCES

- [1] J. Abbott, A. Bigatti, M. Kreuzer and L. Robbiano *Computing Ideals of Points*, J. Symb. Comput. **30**, pp 341–356, (2000).
- [2] J. Abbott, M. Kreuzer and L. Robbiano *Computing zero-dimensional Schemes*, J. Symb. Comput. **39**, pp 31–49, (2005).
- [3] L. Robbiano and J. Abbott (eds.), *Approximate Commutative Algebra*, Text and Monographs in Symbolic Computation, Springer-Verlag Wien, 2009
- [4] D. Bini, M. Capovani and O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli 1988.
- [5] S. Basu, R. Pollack and M.F. Coste-Roy, *Algorithms in Real Algebraic Geometry*, Algorithms and Computation in Mathematics, Vol. 10, Springer-Verlag 2006.
- [6] B. Buchberger, M. Möller, *The construction of multivariate polynomials with preassigned zeros* In J. Calmet Editor, Proceedings of the European Computer Algebra Conference (EU-ROCAM '82, Lecture Notes in Comp. Sci., **144**, Springer, pp 24–31, (1982).
- [7] CoCoATeam, CoCoA: a system for doing Computations in Commutative Algebra. Available at <http://cocoa.dima.unige.it>.
- [8] J. Dégot, *A Condition Number Theorem for Underdetermined Polynomial Systems*, Mathematics of Computation, **70**, n. 233, pp 329–335, (2001).
- [9] D. Eisenbud, *Commutative algebra with a view toward algebraic geometry*, Graduate Texts in Mathematics, Springer, 1995.
- [10] C. Fassino, *Almost Vanishing Polynomials for Sets of Limited Precision Points*, J. Symb. Comput. **45**, pp 19–37, (2010).
- [11] C. Fassino, M. Torrente, *Vanishing Polynomials at Sets of Empirical Points*, submitted.
- [12] L. Gonzalez, H. Lombardi, T. Recio and M.-F. Roy, *Sturm-Habicht sequence*, In Proceedings of ISSAC'1989, ACM New York, USA, pp 136–146
- [13] N.J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 1996.
- [14] M. Kreuzer, H. Poulisse and L. Robbiano, *From Oil Fields to Hilbert Schemes*, in: L. Robbiano and J. Abbott (eds.), *Approximate Commutative Algebra*, Text and Monographs in Symbolic Computation, Springer-Verlag Wien, pp 1–54, (2009).
- [15] M. Kreuzer, L. Robbiano, *Computational Commutative Algebra 1*, Springer, Heidelberg 2000.
- [16] M. Kreuzer, L. Robbiano, *Computational Commutative Algebra 2*, Springer, Heidelberg 2005.
- [17] M. Shub, S. Smale, *Complexity of Bezout's Theorem I: Geometric Aspects*, Journal of the American Mathematical Society, **6** n. 2, pp 459–501, (1993).
- [18] A.J. Sommese, C.W. Wampler, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific, 2005.

DIPARTIMENTO DI MATEMATICA, UNIVERSITÀ DI GENOVA, VIA DODECANESO 35, I-16146 GENOVA, ITALY

*E-mail address:* `robbiano@dima.unige.it`

DIPARTIMENTO DI MATEMATICA, UNIVERSITÀ DI GENOVA, VIA DODECANESO 35, I-16146 GENOVA, ITALY

*E-mail address:* `torrente@dima.unige.it`